Analysing file management behaviour

Jesse David Dinneen
School of Information Studies
McGill University, Montreal
September 2017

A thesis submitted to McGill University
in partial fulfillment of the requirements
of the degree of Doctor of Philosophy

Contents

Figures

Tables

Abstract

In this thesis I examine and improve upon the state of knowledge about a task ubiquitous in and fundamental to the use of computers, file management (FM). This is done in three chapters, constituting a review of the relevant scholarly literature, a methodological contribution, and an empirical study of the FM behaviour of 301 computer users, respectively.

In the first chapter I synthesise hundreds of studies about various aspects of FM to identify a large but previously unacknowledged body of research into FM, and characterise the knowledge and knowledge gaps the literature evinces. I find that studies of FM are typically motivated by understanding users' FM behaviour, the factors that determine this behaviour, and how FM can best be supported by systems and services. After examining the studies' methods and findings, I conclude that the differing goals of past studies have entailed small sample sizes, inconsistent data collection, and incommensurable contexts that preclude deriving a quantitative description of users' behaviour necessary for understanding such behaviour, and that such a description can not currently be derived as no tool exists that collects the necessary data. In the second chapter I describe developing and testing novel data collection software designed to address this knowledge gap and overcome limitations of practicality seen in existing tools, specifically by facilitating collecting extensive data from large and heterogeneous population samples. The software was built with open-source tools and shared with the research community, and can be used and reused in future studies to treat the gaps in knowledge identified in the first chapter.

The third chapter then describes the first use of this software to examine in detail the file systems of 301 participants, producing a broad, quantitative description of typical FM behaviour and facilitating comparison across previously incommensurable studies. The collected data are found to be log-normally distributed, and so the statistical analyses necessary for deriving a meaningful description of such data are made, and indication of this for the accuracy of analyses made in previous studies is discussed. In describing users' behaviour I find that despite the proliferation of alternatives to traditional FM, users are now keeping considerably larger collections than previously observed, and storing these in folder trees that exhibit stable internal structure and file categorisation while becoming taller and wider. These results establish a basic quantitative description of typical FM behaviour, thus laying the necessary foundation for further study of FM, including modelling users and their collections, studying the factors that determine FM behaviour, and advancing theory about the management and organisation of information. The results may also aid in the interpretation of qualitative studies of FM and the design of software and services to support it. Such advances – in knowledge and practical action – are required to better support FM, a daily and fundamental part of using a computer and managing digital content, and are enabled by the problem identification, methodological contribution, and findings of this thesis.

Résumé

Dans la thèse, j'examine et j'améliore l'état de la connaissance concernant une tâche omniprésente et fondamentale à l'utilisation d'un ordinateur : la gestion des fichiers électroniques (GFE). La thèse comprend trois chapitres : un recensement de la littérature scientifique sur le sujet, une contribution méthodologique et une étude empirique du comportement de GFE de 301 utilisateurs d'ordinateurs. Dans le premier chapitre, je synthétise des centaines d'études sur divers aspects de la GFE pour relever un corpus important de recherches encore méconnues sur le sujet, caractériser la connaissance et déterminer quelles sont les lacunes dans la connaissance telles que relevées dans les publications. On y découvre que les études de la GFE sont généralement motivées par la compréhension du comportement des utilisateurs, les facteurs déterminant ce comportement et comment la GFE peut mieux s'appuyer sur les systèmes et les services. Après un examen des méthodes et des résultats des études, je conclus que leurs objectifs différents impliquent des échantillons restreints, des collectes de données inconsistantes et une diversité de contextes qui limitent la capacité des chercheurs à décrire quantitativement le comportement des utilisateurs pour mieux le comprendre et qu'une telle description n'est pas possible à l'heure actuelle puisqu'il n'existe aucun outil pour en recueillir les données nécessaires. Dans le deuxième chapitre, je décris le développement et le test d'un nouveau logiciel de collecte de données conçu pour étudier les lacunes dans la connaissance et surmonter les limites pratiques des outils existants, spécifiquement en facilitant une collecte exhaustive des données à partir d'un vaste échantillon hétérogène issu de divers contextes. Le logiciel a été conçu à partir d'outils en libre accès et partagé avec la communauté de la recherche. Celui-ci peut être utilisé et réutilisé pour d'autres études sur les lacunes dans la connaissance relevées dans le premier chapitre.

Dans le troisième chapitre, je décris la première utilisation du logiciel pour examiner en détail le système de fichiers de 301 participants, permettant ainsi une description générale et quantitative du comportement typique en matière de GFE et facilitant la comparaison entre un nombre incommensurable d'études existantes. Les données recueillies suivent une distribution log-normale et permettent les analyses statistiques requises pour en extraire une description significative. Je discute enfin de l'exactitude de l'analyse des études antérieures. Relativement au comportement des utilisateurs, malgré une prolifération de méthodes de rechange aux solutions traditionnelles en GFE, les utilisateurs conservent maintenant des collections beaucoup plus volumineuses qu'observées auparavant et conservent les fichiers dans une arborescence qui témoigne d'une structure interne et d'une catégorisation des fichiers stables. Par ailleurs, ces structures sont également plus profondes et plus extensives. Les résultats permettent d'établir une description quantitative de base du comportement typique en matière de GFE, posant les fondations nécessaires à de futures études de la GFE, y compris la modélisation des utilisateurs et de leurs collections, l'étude des facteurs déterminant le comportement de GFE, et l'avancement de la théorie de la gestion et de l'organisation de l'information. Les résultats pourront également contribuer à l'interprétation d'études qualitatives de la GFE et à la conception de logiciels et de services pour la soutenir. De telles avancées, à la fois théoriques et pratiques, sont requises pour mieux soutenir la GFE, une tâche quotidienne fondamentale à l'utilisation d'un ordinateur et à la gestion de contenu numérique, et le tout est rendu possible par l'identification de la problématique, la contribution méthodologique et les résultats de la thèse.

Acknowledgments

Research is a social activity and I therefore have many to thank for their respective roles in helping me with this work. First and foremost, I thank my supervisor, Prof. Charles-Antoine Julien, for his open and patient guidance about and support in matters professional, scholastic, financial, and personal, all of which was invaluable to my thesis, scholarly training, and career. Your help has made an impression on me that will last for many years, and I am deeply grateful for it. I also wish to thank the members of my thesis committee, Profs. Ilja Frissen and Jamshid Beheshti, for their essential advice about the direction and preparation of my thesis; they too have always been willing to help me, and I am very appreciative of their help. I thank the whole committee for letting me stand on their shoulders and making this dissertation possible.

I thank Dr. Richard H. Tomlinson for his generous donation to McGill, which resulted in the fellowship that paid for the majority of my doctoral studies and has enabled many people to study when that would have otherwise been impossible. I also thank the McGill University School of Information Studies for their financial support, and the Association for Information Science & Technology and Thomson Reuters for their recognition and funding of my thesis.

I also wish to express my gratitude to many others that have helped in various ways:

- Fabian Odoni for his invaluable assistance in improving the data collection software's data transfer methods, conforming the code I wrote to PEP 8 guidelines, testing the compiled executables, and securely hosting the source code until we deemed it ready to share with the public;

- Christian Brauner and Verena Sebald for sharing their experiences in statistical

analysis and research design, and for aiding in recruitment;

- Jochen Steffens, Daniel Steele, Mohammed AlGhamdi, Banafsheh Asadi, Fei Shu, Kylie Szymesko, Laurie Karnis, the PhD students at the McGill University School of Information Studies, and many others for their camaraderie, support, advice, and assistance in recruitment;

- Dominic Boisvert and Carla DeLancy for their help in identifying bugs in the data collection software;

- Steve Evans for a considerable effort to recruit Linux participants;

- Steve Dodier-Lazaro for his advice and support concerning my recruitment efforts;

- Rob Capra and William Jones for sharing their expertise in personal information management research and advice about my thesis topic;

- Kim Dalkir, Karryn Moffatt, and Catherine Guastavino for their kind and patient advice;

- Jamshid Beheshti, France Bouthillier, Kim Dalkir, Ben Hanrahan, and Shane McIntosh for serving on my defence committee and providing invaluable feedback;

- and Cathy Venetico, Kathryn Hubbard, and Shannon Sullivan for their administrative assistance in all things scholarly.

I also thank the many anonymous others who participated in my study, helped in recruitment by sharing my calls for participation, or helped to develop the many open-source tools I used in carrying out my research and writing about it.

I thank my mother and extended family for their care and encouragement, including especially my late stepfather, Joseph E. Tomkowitz, for fostering me and my curiosity. Finally, I wish to thank my partner, Nicola Vernon, for her joyful and tireless encouragement and support of me in all things, which would require another thesis-length document to adequately describe.

Preface

Contributions in co-authored works

This thesis consists primarily of three papers, constituting Chapters 1, 2, and 3. All three papers are my own work and solely my own writing, but all three benefited from the review, feedback, and advice of my thesis committee, Profs. Charles-Antoine Julien, Ilja Frissen, and Jamshid Beheshti, with Paper 2 receiving additional attention from Profs. Julien and Frissen prior to its submission for publication.

Paper 1 further benefited from comments provided by Prof. Rob Capra at University of North Carolina Chapel Hill, and Paper 2 benefited from comments provided by Fabian Odoni at the Swiss University of Applied Sciences (HTW Chur) and anonymous reviewers for the 2016 ASIS&T annual meeting. I performed all principle programming of the data collection software described in Paper 2, and Fabian Odoni assisted in improving the software's data transfer methods, conforming the existing code to standard style guidelines, testing the compiled executables, and providing hosting of the source code on his university's secure server until we deemed it ready to share on GitHub.

Original scholarship, contributions to knowledge

The following are the elements of this thesis that constitute original scholarship and distinct contributions to knowledge:

- Paper 1

    1. Demarcation and description of existing but previously unlabeled research sub-

field, *file management*, and first explication of its relationship to parent fields

2. Synthesis of relevant, previously disparate studies under umbrella of file management research
3. First identification of motivations, methods, findings, limitations, and future directions evidenced across file management literature, including identification of specific critical knowledge gaps and their relevant necessary methodological improvements

- Paper 2

1. Design and creation of novel data collection tool to treat knowledge gaps and need for methodological improvements identified in Paper 1
2. Empirical validation and refinement of said tool
3. Documenting and sharing said tool's source code with research community

- Paper 3

1. Extensive quantitative description of file management behaviour to treat knowledge gap identified in Paper 1
2. Identification of trends in file management behaviour across previous studies
3. Identification of need for specific statistical analyses for relevant data
4. Design and execution of said analysis

Introduction

File management (FM) is an activity ubiquitous in and fundamental to the use of computers, as files provide a representation of digital contents and folders provide the means to organise and access such files. Supporting users performing FM, through improved software and services, requires understanding relevant phenomena, including users' behaviour, its determining factors, and the systems being used, but despite the ubiquity and importance of FM, knowledge about all of these phenomena is surprisingly limited. The overall goal of this thesis is therefore to understand the limitations of and improve the state of this knowledge.

While a considerable number of prior scholarly works have examined many aspects of FM, they have never been acknowledged as belonging to a single topic of study nor have they been extensively reviewed together, and so their common motivations and methods are unknown and it is unclear what collective knowledge and gaps in knowledge emerge from their synthesis. In the first of the three principal chapters of this thesis I address this by providing such a review, drawing together over 200 publications to demarcate *file management research* and describe the state of knowledge about this topic. Among my findings I identify the need for a broad and confident quantitative description of typical FM behaviour to enable advanced study like modelling users' behaviour and identifying its determining internal and external factors, such as individuals' cognitive differences or conventions encouraged by the operating systems used. I identify that small sample sizes, narrow data collection, and incommensurable contexts have so far prevented such a description from emerging from many previous studies, and note that the existing data collection tools cannot overcome these limitations because they are do not collect the

necessary data and are impractical to administer.

In the second chapter, therefore, I describe developing and testing data collection software to treat these issues. The software, called Cardinal, was built using open-source resources and shared freely on the Internet. It facilitates large-scale, remote, and asynchronous collection of extensive FM behaviour data from anonymous participants by examining 38 file system properties and additional relevant dimensions like demographic, software, and hardware variables. I also describe its use in a 15-day trial implementation and its subsequent revision, and note how it may be used in further FM research.

In the third chapter I use that software to carry out a study that addresses the incomplete quantitative description of users' behaviour identified in the first chapter. In that study I collected file system data from 301 participants, which I then analysed to derive 56 measures of their FM behaviour. The data along most measures were log-normally distributed, and so I used statistical analyses beyond those used for normally-distributed data. The result is an extensive quantitative description of typical FM behaviour, which I discuss and compare piecemeal to the findings of previous studies. The results establish the necessary foundation for further study of FM, including identifying the principal components of users' behaviour, modelling their behaviour, generating a standardised file collection for evaluating new FM systems, studying the factors that determine FM behaviour, and advancing theory about the management and organisation of information. By providing scale to dimensions of behaviour identified in previous studies, the results can also facilitate making targeted improvements to the software and services that support computer users.

In summary, the objectives of the present thesis are:

1. identify and synthesise studies of digital file management, including identifying their common motivations and methods (Chapter 1);

2. describe the state of their collective knowledge, including identifying their findings, limitations, gaps in knowledge, and future directions (Chapter 1);

3. develop software necessary to alleviate a gap in knowledge and the limitations of the

quantitative data collection tools used in previous FM studies, namely: facilitate a broad quantitative description of FM behaviour by collecting extensive data from a large population sample in a practical manner (Chapter 2);

4. collect, analyse, and report on the data necessary to provide an extensive quantitative description of typical file management behaviour to enable further research like those described above (Chapter 3).

While the three chapters of this thesis build linearly upon another and are ordered accordingly, transition sections placed between the chapters are exclusively dedicated to connecting the chapters into a cohesive, single program of research and locating each chapter within that program. A conclusion notes how the thesis's objectives were met and the contributions made, summarises the thesis's findings and limitations, and discusses directions for future research, while the appendices provide additional technical details about the data collection performed.

Chapter 1

The ubiquitous file: a review of digital file management research

Abstract

Computer users spend time every day interacting with digital files and folders, including creating, downloading, naming, moving, saving, copying, reviewing, navigating, searching, and deleting them. This phenomenon of *file management*, a core element of personal information management, has been the focus of many studies across various fields, but has not been explicitly acknowledged or made the focus of dedicated synopsis, synthesis, or reflection. In this paper we present the first dedicated review of this topic and its research, bringing together over 200 publications to examine the common motivations, methods, findings, and future challenges of the field represented by this previously unexamined body of work. The literature evinces three common research motivations: understanding how and why users store, organise, retrieve, and share files and folders, understanding internal and external factors that determine their behaviour, and attempting to improve the user experience through novel interfaces and information services. Several implicit conceptual frameworks, methods of inquiry, and approaches to designing and testing systems are employed in the literature, and open research questions continue to motivate and challenge researchers. It is concluded that file management is a ubiquitous, difficult, relatively unsupported, and not well-understood activity that invites and has received multidisciplinary research with broad importance across information science.

1.1 Introduction

Computer users spend time every day interacting with digital files and folders, including creating, downloading, naming, moving, saving, copying, reviewing, navigating, searching, and deleting them. This ubiquitous activity, called file management (FM), is difficult, personal, deeply psychological in nature (Lansdale, 1988), and so fundamental and common one can hardly imagine a current information professional, IT admin, Web developer, librarian, archivist, modern computer user, or active citizen of the information age who cannot manage files, as it is one of the core activities required for using a computer today for anything beyond casual and lightweight media consumption. FM is also increasingly complex, as improvements in desktop search, the addition of tagging functions to file manager software, and the increasing application of cloud services to FM have expanded the number possible user interactions and challenges, and added further nuance to users' behaviour. Users can keep files locally, synchronise them across devices and in the cloud, organise them as a collection using local and Web-based applications by themselves or in collaboration with others, navigate and search through them in multiple ways, and so on. File management is therefore one of the most central activities involved in using a computer, and thus an important aspect of living in the information society.

FM can be supported by personal information management (PIM) systems, provided their design is informed by an understanding of the behaviour that users exhibit and its determinant factors. Many studies have worked towards improving and implementing this understanding, and yet this literature and their subject have not been formally acknowledged, reviewed, summarised, synthesised, or reflected upon. The goal of this paper is therefore to provide a review of the relevant literature, and in doing so to demarcate the body of scholarly work about file management and understand the current state and limits of its knowledge. In what follows we provide definitions and background, detail the motivations, frameworks, and methods of file management research, outline the relevance of FM research for information science and other fields, and discuss future directions and challenges.

1.1.1 History and context of files

The word *file* can have multiple senses related to computer files (Harper et al., 2013), but the one used in the reviewed literature and adopted here refers to what is perhaps the most common: representations of digital content stored in standard file systems and presented to users through the metaphor of a paper file (e.g., a physical document, not to be confused with the British sense of file *as a folder*). In simple cases, files are used to represent, for example, a document, an image, some audio data, an executable program, a database, an ongoing session in some software, or an archive of any of such files. Folders extend the file metaphor to provide categorised access to files and to more folders by containing them, and are presented to the user as though arranged in a spatial hierarchy that starts at a common *root* folder and may contain a minimal default folder structure. Though the term *directory* is often used interchangeably for *folder* (and the preferred term for users of some operating systems), in this thesis we use *folder* when referring to the user's experience or view of such an item (e.g., what they see in their file manager, what they move and rename) and *directory* when referring to applications' view of or interaction with locations within the file system.

Users of all contemporary operating systems (OSes; Windows, Mac OS, GNU/Linux, BSD, Solaris, Android, iOS, Windows Mobile...) interact with files. The user creates, names, renames, downloads, uploads, attaches, copies, organises, cuts, pastes, tags, links via symlink or shortcut, navigates, searches, deletes, and restores files and folders while using a computer, in contexts that may be occupational (e.g., personally managing company files) and personal (e.g., maintaining files for personal use). The focus of this review is research into how users interact with digital files presented *as files*; we therefore define file management as any user activity involving the actions listed above, though additional relevant actions are conceivable and may become common in the future. In other words, our definition includes what is seen in a file manager or applications' file open and save dialogues, but not items in other contexts that may have items resembling files or allow interactions similar to those of a file manager; for example, emails downloaded to a laptop and saved as eml files into a folder are indeed files, but those emails presented to a user

in a Web-based mail client – possibly sortable into folders – are not considered files for the purposes of this thesis because they are not accessed from or acted upon from within the file manager. Digital items may also be viewed and managed with items of the same format in particular applications, for example as a collection of songs in iTunes or photos in a photo viewer. Although these items likely also exist as files, we do not regard managing the items within the format-specific application as file management. The relevance of such contexts for FM is addressed later in this chapter.

The metaphor for digital content as files organised in folders has historical roots as far back as the 1960's (Corbató, Merwin-Daggett, & Daley, 1962) and has been pervasive in computing for over 40 years (Harper et al., 2013). Though this metaphor for digital content has been questioned (Halasz & Moran, 1982) and warrants critical reflection and refinement (Harper et al., 2013), it is one of the oldest in computing, is widely used, and is currently without a serious alternative. Operating systems store, handle, represent, and manage files and folders differently than how they are presented to and handled and managed by the user (Harter, Dragga, Vaughn, Arpaci-Dusseau, & Arpaci-Dusseau, 2012); for example, in POSIX systems (Mac OS, GNU/Linux), folders (directories) are actually files, devices are represented as files, and the OS and its applications may read and write to a file many times while the user is simply viewing it. This review focuses primarily on files and folders as they are presented to the user, but user-file interaction is of concern as much to those designing file systems as it is to those seeking to understand users' behaviour and improve the relevant software and interfaces.

The original method for performing file management was to enter commands, like *mv* for moving and *cp* for copying, into a command line prompt. This method persists today, though given the popularity of graphical desktop environments it is likely that most file management is done in graphical file manager applications and dialogues initiated, for example, when opening a file in an application, to directly manipulate file and folder icons. In Microsoft and Apple's desktop OSes, graphical software for managing files is provided by default (File Explorer and Finder, respectively); many alternative file managers are available, each with different features and views of files, but it is unclear if most users

install or are even aware of these. When using Linux, users may or may not be given a default graphical manager by their distribution, and may install alternative terminal-based or graphical file managers. Regardless of the operating system, users likely spend much time performing the actions described above; so far as we know the exact time spent managing files per day or year has never been calculated for an individual or collectively, but given that the most recent estimate from the US Census Bureau is that 78.5% of all households have at least one desktop or computer (File & Ryan, 2014), it is reasonable to assume the aggregate time spent interacting with files is considerable.

Though the antecedents of FM research can be found as early as the 1980's, for example in studies of how people manage paper documents (Case, 1986; I. Cole, 1982; Malone, 1983), FM has not been explicitly acknowledged as a topic of study. This is despite the relatively large amount of attention the phenomenon has received in research and despite connections to and shared interests with other fields of study, described in this paper. We next describe our methodology for reviewing the relevant literature, and then proceed to the review.

1.1.2 Review methodology

The goal of this review is to demarcate the body of scholarly work about the management of digital files and folders in common computing environments (i.e., desktops, laptops, tablets, and mobile phones) and understand the current state of knowledge about the immediately relevant phenomena identified in such work. The specific topic of interest we therefore sought to collect and review literature about was people's interactions with files or folders. We did this identifying and searching scholarly research databases (e.g., Web of Science, Google Scholar) that index journal articles and conference proceedings dealing with, for example, personal information management, human-computer interaction (e.g., proceedings of ACM conferences), interface design, information behaviour, information science (e.g., *Information Research* and *JASIST*), computer software development, and personal digital archiving. We searched with keywords including *personal information management*, *file management*, *file system*, *desktop management*, *folder organisation*, and

file retrieval and various additional permutations. We then scanned the manuscripts' references to identify additional relevant articles and proceedings (i.e., citation pearl growing; Ramer, 2005). We filtered out manuscripts describing information management, personal or otherwise, or general computer use, unless file management, presentation, or similar concepts were primary topics of the works. We did include in our review, however, additional tangential works if they commented on or helped to elucidate trends or topics seen elsewhere in the literature.

The result of our literature search was 211 manuscripts with publications dates from 1960 to 2016, including reports of quantitative and qualitative empirical studies, the development of novel systems at various stages of completion, opinion pieces (e.g., reflection on interface design), and reviews (e.g., about PIM; to our knowledge, no review of FM exists). We analysed these manuscripts to capture common themes, such as motivations and findings, concepts and methods, limitations, and directions identified for future research. The results of our analyses are discussed in turn.

1.2 Motivations of file management research

In this section we present FM literature along three common motivations: understanding user behaviour (or *what* users do), understanding the individual differences and external factors in this behaviour (or *why* they do it), and aiming to improve file management systems and software (or *how* to better support users and their behaviour). Each is discussed and presented with a table summarising the relevant literature, and the theoretical and conceptual frameworks and methodologies employed across the studies are examined in the next section.

1.2.1 Understanding user behaviour

Many studies seek to understand users' FM behaviour, albeit under different topic banners like personal information management, personal digital document management, and personal digital archiving. Categorisation of this literature reveals four common themes

in studies motivated to understand FM behaviour, although works rarely fall into just one category: file and folder storage (e.g., creating, downloading, naming, managing backups), organisation (e.g., organising the folder structure and categorising files into it), retrieval (e.g., searching, navigating, and tagging), and sharing (e.g., managing shared folders, sending files). The first three of these categories are loosely synonymous with keeping, meta-level, and refinding activities (Jones, 2007b), also known as keeping, organising, and exploiting activities (Whittaker, 2011), while sharing activities entail and happen across all three of the other activities. The literature reviewed is reported here along these themes, each entailing characterising users' behaviour (or the outcomes of their behaviour)[1] and the challenges users face in performing relevant actions. Discussion of behaviour *beyond* the FM context but still within the purview of PIM research can be found in a recent encyclopaedia article on PIM (Jones, Dinneen, Capra, Pérez-Quiñones, & Diekema, 2015).

Storing

Actions done to store files and folders include creating, downloading, naming, moving, copying, backing up, and synchronising (or syncing). Though reports of the number of files users store vary greatly, the number is always large: recent studies have found averages from approximately 4,700 (Hicks, Dong, Palmer, & McAlpine, 2008) to 15,000 files per user (Massey, TenBrook, Tatum, & Whittaker, 2014), with minimums as low as 1,000 (Gonçalves & Jorge, 2003b) and maximums as high as 56,994 (Whitham & Cruickshank, 2017). The number of folders stored also varies across studies, for example from 56 (Boardman & Sasse, 2004) to 1,044 (Henderson, 2005), and in a study of one organisation the average increased from 2,400 to 8,900 (i.e., a 370% increase) over a five year period (Agrawal, Bolosky, Douceur, & Lorch, 2007).

Users' files come from various sources, including the Web (Jones, Bruce, & Dumais, 2001; Huvila, Eriksen, Häusner, & Jansson, 2014), external devices (Capra, Vardell, &

[1]The distinction between behaviour and the outcomes of behaviour is not always a clear one and is often not explicated in FM research. Doing so remains a challenge, and no attempt to solve it is made in this thesis. Rather, the two concepts are used interchangeably when it is not obviously a problem to do so.

Brennan, 2014), and peer-to-peer or cloud software (Marshall & Tang, 2012), though do not often come from their cell phones, despite the ability to download files to smart phones from the Web (Capra, 2009). Several studies have sought to understand the contents of users' collections, finding for example that document and image files are the most common types kept by students and knowledge workers (Gonçalves & Jorge, 2003b; Hicks et al., 2008), and that files may be regarded by users as ephemeral, archived, or current for their intended use (Nardi, Anderson, & Erickson, 1995).

Understanding the challenges of storing so many files is a primary concern of FM studies, and several challenges have been identified. Some challenges are due to the imperfect analogue between the digital desktop and its files with the physical counterparts they are modelled after. For example, some users do not understand the desktop's location in relation to the rest of the accessible disk (Ravasio, Schär, & Krueger, 2004), and for some it is not an attractive place to store files since it is often covered with other windows and does not have the multiple flexible views of its content that the file manager provides (Kaptelinin, 1996). Other challenges are due to the proliferation of digital files: with so many files stored, it is difficult to remember that a file exists in time to use it when it is needed (Jones, Dumais, & Bruce, 2002).

File and folder naming behaviour is one concern related to storage habits, as generating meaningful, descriptive but concise, and unique names for files and folders also poses a challenge. In studying file naming behaviour, users have been found to exhibit considerable creativity in file naming (Carroll, 1982), though patterns are identifiable: files are named to display the document they represent, their purpose, a project title, or a date (Hicks et al., 2008). Folder names have been found to represent their files' genre, a relevant task, a particular topic meaningful to the user, or a period of time (Henderson, 2005), they may also represent a priority ranking, their use as storage, or a combination of these and other themes (Khoo et al., 2007). Beyond alphabetical characters, users also make use of numbers and punctuation such as white space, the underscore, and the hyphen (Gonçalves & Jorge, 2003b). The mean length of users' file names may be increasing as the system's limits increase: studies have shown an increase from 6 characters (Carroll,

1982), to 12.6 (Gonçalves & Jorge, 2003b), and recently to 18.8 (Fitchett & Cockburn, 2015). Despite all the creativity and possibilities in file naming, duplicated file and folder names are common (Henderson, 2005; Hicks et al., 2008) and are further increased by system-generated folders (Henderson & Srinivasan, 2009); this poses an obvious challenge to retrieving files whether by navigating or searching.

The introduction of the cloud and desktop synchronising software, such as Dropbox, has likely changed the nature of users' file storage behaviour, though the nature of this change is still being investigated. Users are confused by the cloud and by syncing software: they do not understand what such software is, does, or how it interacts with their local storage or other cloud software (Marshall, Wobber, Ramasubramanian, & Terry, 2012). They may conceive of it as a file repository, shared repository, personal replication store, shared replication store, and synchronisation mechanism (Marshall & Tang, 2012), and try to understand it as it relates to their local storage (Tang, Brubaker, & Marshall, 2013). Users' storage behaviour on the cloud requires further study, as we discuss again below in the context of file sharing behaviour.

A growing portion of studies have examined how and when users back up their files and folders. Though what exactly constitutes a back up is conceived of variedly in the literature, it typically refers to copies of valuable portions of a collection made to provide redundancy or version control, stored on separate physical media of various formats, and not frequently accessed or modified. People may rely on dedicated back up, sharing, or syncing services such as Dropbox or Apple's Time Capsule to make their back ups (Marshall & Tang, 2012), but also may not feel these are reliable in their schedules or operations, and so may initate and make back ups manually (Dearman & Pierce, 2008; Capra et al., 2014).

Organising

Organising actions include renaming, creating subfolders, creating shortcuts, symlinks or hard links, filing (e.g. downloading and then moving), copying directly or by pasting, moving directly or by cut and paste, and deleting. This is typically done using the

folder hierarchy, and for various reasons, including giving the user a place for files to persist (Whitham & Cruickshank, 2017) and to help them make sense of, summarise, group, and maintain an overview of the files (Jones, Phuwanartnurak, Gill, & Bruce, 2005; Ravasio et al., 2004; Whitham & Cruickshank, 2017), which in turn aids memory about the files organised so that they may be retrieved later (Whitham & Cruickshank, 2017; Xie, Sonnenwald, & Fulton, 2015). Popular file manager software aims to facilitate this process, but with uncertain results: some users have reported that the locations of OS-provided default folders are confusing and that the system-ascribed metadata was not useful for understanding their collections (Ravasio et al., 2004), while others make distinct use of the desktop, default folders, and secondary folders (Paré, 2011). Despite this, a few studies have found that users do indeed store files in default folders, including files that are *active* or currently being frequently accessed (Bergman, Whittaker, Sanderson, Nachmias, & Ramamoorthy, 2010), in locations such as *My Documents* and on the desktop (Khoo et al., 2007), and that use of the default folders among users at one organisation grew over five years, accounting for 40% of all files (Agrawal et al., 2007).

Users also create, arrange, and remove subfolders in other locations, such as the root of their hard drives (Ravasio et al., 2004) or in their respective home folders, and in doing so they determine the shape of the overall folder tree with which they interact. By studying properties of the tree, like its height or depth (the maximum number of steps taken when navigating into consecutive subfolders) and consistency (deviation in shape among its main branches), a quantitative description of how people organise their digital items can be provided and from this particular aspects of user behaviour can be determined and described. Descriptions such as these have been provided in many studies, including in those using quantitative data to complement and give scale to their qualitative findings. Among disparate contexts, participant groups, and file system measures used in previous studies, users' organising behaviour has been found to vary wildly.

As mentioned above, users may be spreading their files across as few as 56 folders or as many as 9000. Figures reflecting the total number of folders are of limited use in analysing the organisation behaviour, however, as folders can contain any number of files;

they can, for example, contain many files, acting as traditional storage locations, or only other folders, thus acting as a navigation fork (Bergman et al., 2010).

User-created hierarchies may vary greatly in maximum depth; a range from one level of depth (i.e., no subfolders) to sixteen levels deep was found in a single study (Henderson & Srinivasan, 2011). Deeper structures are in part a result of larger collections (Henderson & Srinivasan, 2009), and depth in turn contributes to an increase in file name redundancy (Henderson, 2011) and time required to retrieve files (Bergman et al., 2010). Users may create hierarchies that display consistency among their internal *branches* (Gonçalves & Jorge, 2003b) or not (Henderson, 2005). Hierarchies may present the user with many navigation decisions by having a high average *branching factor*, or number of subfolders per folder, of 41.8 (Hicks et al., 2008), or a very low branching factor, for example of only 1.84 (Gonçalves & Jorge, 2003b).

The location of files within the folder hierarchy is another object of inquiry in FM studies, as where users put their files later affects how long it takes to retrieve them (Bergman, Gradovitch, Bar-Ilan, & Beyth-Marom, 2013a). Users may file every single document despite each classification action being cognitively demanding (Ravasio et al., 2004), or may leave up to 6.5% unfiled (Henderson & Srinivasan, 2011). Depending on the user, filing the average file may mean storing it just two levels down from the root of the tree (Bergman, Whittaker, & Falk, 2014), while others store most files in deeper levels (Hicks et al., 2008). As the number of files in a folder increases, so does the work required to review them all, and although users report creating new subfolders when a folder contains 3-7 items (Ravasio et al., 2004), the average number of files found in folders has ranged from low figures like 0 (Henderson & Srinivasan, 2009) or 4 (Zhang & Hu, 2014) to 12 (Bergman et al., 2010; Henderson & Srinivasan, 2009) or 16 (Gonçalves & Jorge, 2003b; Hardof-Jaffe, Hershkovitz, Abu-Kishk, Bergman, & Nachmias, 2009b). Default folders have been found to have a mean of 19.42 files per folder (Bergman et al., 2010), and so may be fuller than folders in completely user-created branches, perhaps because they are more likely to be used to store frequently accessed documents; a complete comparison will require examining and comparing folder structures beyond those housing

recently accessed files, including on devices where backups or personal archives are stored.

Studies have conflicting reports of users creating folders without putting files into them: while one study found users typically do not create empty folders (Khoo et al., 2007), another found that most users do, with 8% (mean) of folders being empty (Henderson & Srinivasan, 2009), and higher percentages being reported in other studies, including 14% (Sienknecht, Friedrich, Martinka, & Friedenbach, 1994) to 18% (Douceur & Bolosky, 1999). The contexts for such studies vary, however, from university employees to employees in a single corporation. Thus far, what users seem to have in common is limited to their lack of reliance on soft file linking features such as aliases in Mac, shortcuts in Windows, and symlinks in Linux (Gonçalves & Jorge, 2003b; Ravasio et al., 2004). Among the varied findings, different approaches or strategies to organising have been identified, albeit rather broadly, so that we can describe organisers as: neat or messy (Boardman & Sasse, 2004), prone to saving or deleting (Berlin, Jeffries, O'Day, Paepcke, & Wharton, 1993), and prone to filing or piling (Malone, 1983), extensive filing or single folder filing (Henderson & Srinivasan, 2011), or mixing approaches (Trullemans & Signer, 2014a). To draw conclusions beyond these, studies are needed with commensurable contexts, participant characteristics, file system measures, and results reporting (Dinneen, Odoni, Frissen, & Julien, 2016).

Retrieving

Retrieving files and folders may be done to find them for the first time (e.g. in a shared drive) or to refind them, which is distinct from simply finding them again because the user has additional information about their existence and location and thus may have additional retrieval methods available (Capra, Pinney, & Perez-Quinones, 2005). Specifically, retrieving can be done manually, for example by navigating through the folder hierarchy to a file's location, or by searching, for example by file property, keyword, or tag label. Both approaches to retrieval require remembering *something* about the object to be retrieved: its location, name, or other properties.

Much FM research has been motivated by understanding navigation and comparing it

with search, typically by examining users' behaviour and preferences and their influences. A preference for navigating to files is much more common than a preference for searching (Fitchett & Cockburn, 2015), even among users who prefer to search rather than navigate folders when retrieving their emails (Jones, Wenning, & Bruce, 2014). There are numerous potential causes for this; users report that they feel desktop search tools are too complicated (Ravasio et al., 2004), the search results are too numerous and not meaningfully ranked (Fitchett & Cockburn, 2015), and that navigating through folders provides important reminding cues about their collections (Barreau & Nardi, 1995).

These reports are reflected in users' behaviour: users perform navigation far more then searching (Bergman, Beyth-Marom, Nachmias, Gradovitch, & Whittaker, 2008; Fitchett & Cockburn, 2015), even when they knew the name of what they were looking for (Teevan, Alvarado, Ackerman, & Karger, 2004), are given improved search engines (Bergman, Beyth-Marom, Nachmias, Gradovitch, & Whittaker, 2008), or have not made the effort to maintain a highly-structured information organisation (Teevan et al., 2004). Users search their files only as a last resort, when navigation fails (Bergman, Beyth-Marom, Nachmias, Gradovitch, & Whittaker, 2008; Fitchett & Cockburn, 2015; Nardi et al., 1995). This is likely in part due to navigation being easier to perform: it allows users to explicate less of their information need and the folders presented at each step provide additional context to guide the navigation (Teevan et al., 2004). This explanation has been given additional weight by two recent studies: one found that navigating tasks required less cognitive effort of participants than searching tasks did (Bergman, Tene-Rubinstein, & Shalom, 2013), while another found that large portions of the brain dedicated to spatial cognition and used in real world navigation are activated during FM navigation, whereas the smaller areas dedicated to linguistic processing were activated during search tasks (Benn et al., 2015).

A third option in retrieving files is to search by tags; tagging provides an alternative to classifying files into folders by allowing users to assign numerous labels to files that can later be searched or browsed and taxing classification and navigation tasks can be avoided, for example deciding which single folder a file should be placed within. Because

of its promise and use in Web-based contexts and email, tagging has been studied in contexts beyond FM, where users have been found to be less preferred than hierarchical navigation (Civan, Jones, Klasnja, & Bruce, 2008) and to entail cognitive load of its own in deriving label names (Gao, 2011). This has been reflected in FM research into tagging, where users report being less frustrated with folders than tags, and in the end use folders more than tags (Bergman, Gradovitch, et al., 2013a) even when their reported preference was for tagging and they were provided both systems (Bergman, Gradovitch, Bar-Ilan, & Beyth-Marom, 2013b). Experienced users may tag faster than they file (Voit, Andrews, & Slany, 2012b), but rarely apply more than one tag (Bergman, Gradovitch, et al., 2013a), thus losing some of the value of the potential for multiple classification of files. The takeaway is somewhat unclear, as noted by Bergman, Gradovitch, et al. (2013b): from the findings of many studies of tagging, one can see that both folders and tags are better, worse, and no different than their alternative at any given aspect of retrieval. Therefore, work remains to provide the kind of explanation for tagging-vs-filing behaviour and preferences that has recently been done for searching-vs-filing.

Despite relatively clear indications that users prefer and perform navigation in FM contexts, desktop search has a discrete purpose and tagging shows promise, and therefore search and tagging systems will likely continue to develop in the coming years. Improved FM tools may, for example, usefully integrate search and navigation functions (Julien, Asadi, Dinneen, & Shu, 2016), or improve searching capabilities by utilising the extensive metadata that users are more likely to remember (Gonçalves & Jorge, 2008a), such as file provenance (Jensen et al., 2010), file type (Blanc-Brude & Scapin, 2007), and time (Dumais et al., 2003).

Sharing

Interacting with shared files and folders typically involves sharing them or having them be shared with you, and then performing the typical storage, organisation and retrieval tasks in a way influenced by the fact that they are shared. Sharing files may be a relatively simple and singular act, for example when users share files on USB sticks or in email

attachments for personal purposes (Capra et al., 2014) or to circumvent institutional access control policies or difficult software interactions (M. L. Johnson, Bellovin, Reeder, & Schechter, 2009). It may, however, be a complex negotiation of a shared information space in a Dropbox folder or company intranet (Tang et al., 2013), or a combination of services that leaves the files fragmented across multiple locations (A. Voida, Olson, & Olson, 2013).

Various problems arise in shared file management contexts. Individual information access strategies break down when managing group information because people struggle to find files in information structures created by others for their own use (Berlin et al., 1993). This problem may be called a lack of *mutual intelligibility*: customisation to make information structures more meaningful for one person often makes them less accessible or intelligible to others (Dourish, Lamping, & Rodden, 1999). If this is treated with an inclusive approach where nothing is deleted, files become forked across multiple versions, folders may get messy, and some users may run out of hard drive space (Capra et al., 2014). If, however, users intend to tidy the shared space, it may be unclear to them who owns any given file (Zhang & Twidale, 2012), will typically face a lack of policy regarding deletion, naming, and organisation (Capra et al., 2014), and will perform moving or deletion actions that other users may later be frustrated to be unaware of (Zhang & Twidale, 2012). In turn, retrieval in shared folders is more time-consuming and prone to error than retrieval from ones own folders, and users may prefer the simple sharing acts to co-managing a shared information space (Bergman, Whittaker, & Falk, 2014). These problems may lead to the establishment of conventions for behaviour involving the shared space, but users report that these are difficult to establish and follow (Mark & Prinz, 1997), although they may in turn be useful, for example for establishing a division of labour involving the files' contents (Wulf, 1997). Therefore, implicit and assumed rules often guide users' behaviour (Zhang & Twidale, 2012), and these clearly warrant further study.

Table 1.1 presents studies that have examined user behaviour, categorised by FM behaviour theme.

FM theme	Example actions	Studies
Storing	creating, downloading, filing, naming, backing up files	Barreau (1995); Capra (2009); Capra et al. (2014); Carroll (1982); Dearman and Pierce (2008); Gonçalves and Jorge (2003b); Henderson (2005); Henderson and Srinivasan (2009); Hicks et al. (2008); Huvila et al. (2014); Jones et al. (2001, 2002); Kaptelinin (1996); Khoo et al. (2007); Marshall et al. (2012); Marshall and Tang (2012); Nardi et al. (1995); Ravasio et al. (2004); Tang et al. (2013)
Organising	creating subfolders, moving and deleting files and folders	Boardman and Sasse (2004); Bergman et al. (2010); Berlin et al. (1993); Henderson and Srinivasan (2009); Henderson (2011); Henderson and Srinivasan (2011); Hicks et al. (2008); Kaptelinin (1996); Malone (1983); Paré (2011); Trullemans and Signer (2014a); Gonçalves and Jorge (2003b); Hardof-Jaffe et al. (2009b); Henderson (2005); Jones et al. (2005); Ravasio et al. (2004); Whitham and Cruickshank (2017); Zhang and Hu (2014)
Retrieving	navigating, searching, tagging files and folders	Barreau and Nardi (1995); Benn et al. (2015); Bergman, Beyth-Marom, Nachmias, Gradovitch, and Whittaker (2008); Bergman, Gradovitch, et al. (2013a, 2013b); Bergman, Tene-Rubinstein, and Shalom (2013); Bergman, Whittaker, and Falk (2014); Cutrell (2006); Cutrell, Dumais, and Teevan (2006); Fitchett and Cockburn (2015); Jensen et al. (2010); Jones et al. (2014); Nardi et al. (1995); Ravasio et al. (2004); Teevan et al. (2004); Voit et al. (2012b)
Sharing	sending files, negotiating storage, organisation, retrieval in shared space	Berlin et al. (1993); Bergman, Whittaker, and Falk (2014); Capra et al. (2014); Dourish, Lamping, and Rodden (1999); M. L. Johnson et al. (2009); Mark and Prinz (1997); Tang et al. (2013); A. Voida et al. (2013); Wulf (1997); Zhang and Twidale (2012)

Table 1.1 – FM studies seeking to understand user behaviour, presented along common themes

1.2.2 Understanding individual differences and external factors

Understanding user behaviour and supporting it with improved software both entail understanding how users' individual differences and broader contexts could determine their behaviour. The few studies of these factors' roles in FM are discussed below; a review of their roles in standard PIM contexts like email, the Web, and paper documents is provided by Gwizdka and Chignell (2007).

Individual differences

Though it is acknowledged that PIM is deeply personal and psychological (Lansdale, 1988), the current state of knowledge about how individual differences affect users' behaviour is still minimal, especially with regards to FM.

Most of the concern for individual differences in FM contexts has been on spatial cognition, which is reasonable given that the only file and folder metaphor in use today is spatial: folders are represented as being contained within one another and displayed in a two-dimensional space, and users *navigate* through the folder hierarchy. An early study of FM found that participants with low spatial ability took twice as long to complete navigation tasks in terminal-based (i.e. text-only, without icons) hierarchical file systems

(Vicente, Hayes, & Williges, 1987), although this difference could be partially alleviated with the addition of a simple map (Vicente & Williges, 1988). The terminal-based paradigm for file interaction is no longer the predominant one, and as of yet no work has specifically looked for similar effects in the modern graphical paradigm. It has, however, been noted that users do develop preferences towards using either the spatial layout of their folders or patterns in file names when retrieving their files (Krishnan & Jones, 2005), suggesting an active role of spatial ability in modern FM. As discussed above, some physiological evidence has recently shed light on why spatial cognition plays such a role (Benn et al., 2015), and so future studies may carry out finer-grained investigations of how different FM actions are affected by this role.

An alternate direction for studying individual differences in FM is the role of personality style: a recent study extended work on the influence of personality on the organisation of physical spaces (Gosling, Ko, Mannarelli, & Morris, 2002) into the realm of digital files by examining cues that participants assumed would predict FM organisation characteristics (Massey et al., 2014). Conscientious participants were found to keep fewer files overall, more files per folder, more files on the desktop, and be more active organisers. Surprisingly, neuroticism and openness were not correlated with organisational or storage behaviour; further work with additional measures of FM behaviour would facilitate a deeper understanding of how personality affects FM behaviour and can be effectively supported, for example through detailed user modelling.

External factors

It is established that external or contextual factors such as occupation, information task, or time are important to understanding the use of paper documents (Kwasnik, 1991) and digital PIM systems (Capra & Perez-Quinones, 2006). This is a concern in FM research as well, but these factors are not yet well understood. For example, the specific effects of occupation are unclear: though participants' occupations have been suggested to be a factor in determining folder naming strategies (Khoo et al., 2007), folder tree height (Zhang & Hu, 2014), and folder organisation (Paré, 2011), occupation seems to have

no effect on branching factor (Gonçalves & Jorge, 2003b), and findings disagree about the effect of occupation on the total number of files stored (Gonçalves & Jorge, 2003b; Agrawal et al., 2007; Henderson & Srinivasan, 2009).

The specific effects of occupation may become clearer as they are explored more narrowly. This may include specific occupational traits like regular activities, demands, and patterns and constraints on time spent organising and retrieving information. Notably, the personal or collaborative management of work files is likely determined in part by institutional policy, for example to delete anything older than two years, or keep everything for at least five years; such policy may be followed, thus determining the contents of a file collection as they do with email (M. L. Johnson et al., 2009), or circumvented if employees find them too onerous (M. L. Johnson et al., 2009).

Another external factor of concern in FM research is the tools used to perform FM: the PIM tool adopted for some task enables, restricts, and affects behaviour of the user (Boardman & Sasse, 2004; Fertig, Freeman, & Gelernter, 1996b), as do tools' surrounding software environments (Kaptelinin, 1996) and the hardware they are housed in. In the context of FM, this includes the computer or hardware device, hard disks, file manager software (sometimes called a file browser – the most popular of which are File Explorer in Windows and Finder in Mac OS), windowing environment (if any), and the operating system (OS).

Though the exact differences between the software relevant to FM have yet to be thoroughly catalogued – for example, the differing OSes and their respective file manager applications allow, encourage, discourage, and forbid different interactions with files – it has been suggested by ancillary analyses in several studies (Agrawal et al., 2007; Barreau & Nardi, 1995; Massey et al., 2014) that such differences may affect users' file storage, management, retrieval, and sharing behaviour. For example, an additional finding of Massey et al. (2014) was that among participants using Windows, conscientiousness was positively correlated with the number of files kept on the desktop, but no such correlation existed for Mac users.

Only one study has explicitly investigated such potential affects (Bergman, Whittaker,

Sanderson, Nachmias, & Ramamoorthy, 2012): while participants retrieved files from their own computers, the researchers noted participants' operating system, file manager presentation mode, retrieval times and success rates, and file and folder organisation. Though collecting data only about recently accessed files, they found that Mac users retrieve files faster than Windows users as a result of a differing organisational strategy: they keep more folders close to the root, with fewer files but more subfolders per folder. They also found that the file manager presentation style with which users performed retrievals best was the icons view, regardless of the OS, and therefore suggested that the Windows default should therefore be changed from the details-based view; users rarely change such defaults (Barreau, 1995). This constitutes a good starting point for understanding the effect of the tool on FM behaviour, and future studies may therefore seek to understand the effects of the OS, file manager, and cloud storage software on storage behaviour and additional variables in organisational behaviour exhibited across participants' recent and archived files.

Hardware, too, may affect users' FM behaviour; limited available hard drive space may cause users to save fewer large files or transfer files to the cloud or external physical drives, and users may be less likely to perform intensive FM actions (like navigating deep trees or making backups) when using a laptop (i.e., using a touchpad, relatively small monitor, and small hard drive) than they would be with typical desktop hardware. Few FM studies have touched upon such topics, but the growth of hard drive capacity, and thus of file storage capacity, can be seen over time in the FM literature. For example, in the mid 1990's users had, roughly, only 80 MB to 1.5 GB of storage space (Nardi et al., 1995), but in a study of one work place taking place a decade later, the mean capacity per participant increased from 8 to 46 GB over a five year period (Agrawal et al., 2007). In that study mean hard drive consumption grew from only 3 to 18 GB across five years, suggesting that at least the employees at that organisation are not restricted by hard drive space, but the adoption of faster, smaller solid state hard drives may introduce another factor into this trend.

A table summarising the individual differences and external factors that have been

examined for their role in determining FM behaviour are presented in Table 1.2.

Group	Factors	Relevant literature
Individual differences	personality style, spatial cognition and ability, perceived importance of documents	Benn et al. (2015); Kwasnik (1991); Lansdale (1988); Massey et al. (2014); Paré (2007); Vicente et al. (1987); Vicente and Williges (1988)
External factors	tool (hardware, OS, FM software), context, information type, time, occupation, task	Agrawal et al. (2007); Barreau (1995); Bergman et al. (2010, 2012); Douceur and Bolosky (1999); Fertig et al. (1996b); Gonçalves and Jorge (2003b); Henderson and Srinivasan (2009); Jones et al. (2002); Kaptelinin (1996); Khoo et al. (2007); Nardi et al. (1995); Paré (2011); Zhang and Hu (2014)

Table 1.2 – FM studies seeking to understand individual differences and external factors determining FM behaviour

1.2.3 Improving FM systems

One of the main goals of FM research, as with broader PIM research, is to save users time and effort, and to understand and support their behaviour through improved file management software.[2] There have been many attempts at this, generally either in the form of augmentations to existing FM software or new and alternative metaphors for handling digital content intended to replace some or all of the hierarchical arrangement of files and folders. In both cases the systems are generally purposefully designed, prototypes are built, and these are then tested with live users in semi-natural use or structured experiments (such methodologies are reviewed later in this paper). Although these systems typically have short lives and do not transfer into mainstream use, the novel concepts they develop and evaluate often do eventually trickle into commonly used software (Kljun, Mariani, & Dix, 2015b). We review here both augmentations to FM existing software and alternative approaches to managing personal digital content.

FM software augmentations

One approach to facilitating file management is to design augmentations to existing FM software to test intuitions about improvements in FM interaction and treat challenges

[2]This kind of software, intended for users of a wide variety to manage and organise files and folders as described in this article, should not be confused with the once similarly named file management or file processing systems of the 1980's, which were essentially simplified database management systems(Hecht, 1985)

identified in previous studies. By aiming to incrementally improve the current file management paradigm this approach benefits from not overloading users with the task of learning a new system (Bondarenko & Janssen, 2005) or surprising them with unfamiliar metaphors or interfaces (Seebach, 2001).

One motivation in augmenting the file manager is to aid the user when navigating through the folder hierarchy. The oldest of these augmentations improved navigating the folder hierarchy in the command line by providing a map of the hierarchy with the user's current location (Vicente & Williges, 1988), and this was found to enable users with low spatial ability to perform retrieval tasks with the same efficacy as users with high spatial ability. More recent attempts improve graphical navigation, for example by highlighting a path to folders that contain file search matches (Fitchett, Cockburn, & Gutwin, 2013, 2014). Navigation has also been improved by allowing users to de-emphasise files (Bergman, Tucker, Beyth-Marom, Cutrell, & Whittaker, 2009; Bergman, Elyada, Dvir, Vaitzman, & Ami, 2014) and by hiding unused folders (Lee & Bederson, 2003) so that fewer navigation decisions are required during re-finding tasks.

As discussed above, there are instances where re-finding by navigation fails and desktop search may be used as a last resort; several studies have therefore sought to augment the relevant software used in such cases. Most of these have focused on improving general search algorithms and interfaces (B. Cole, 2005; Kim & Croft, 2010; Ghorashi & Jensen, 2012; Sauermann, Bernardi, & Dengel, 2005) or applying semantic search to the desktop (Adrian, Klinkigt, Maus, & Dengel, 2009; Handschuh, Möller, & Groza, 2007; Sauermann et al., 2006), for example by using semantic attributes to enhance search ranking (Chirita, Costache, Nejdl, & Paiu, 2006). Others, however, have sought to support specific search contexts, such as finding similar or duplicate files (Manber, 1994) or supporting search with a more interactive interface and drawing on a detailed file metadata index (Liu & Feng, 2016). Further tools for searching across PIM objects beyond files and folders are reviewed by Cutrell (2006).

Several FM software augmentations have been motivated by improving the social and networked aspect of file management by supporting the management of shared and cloud-

based files and folders. Some augmentations simplify the users' interactions, for example by providing a unified view of local and cloud folders (Jones, Thorsteinson, Thepvongsa, & Garrett, 2016), using content and task analysis to suggest locations for new documents to be placed (Prinz & Zaman, 2005), or unifying synchronisation across a users' devices and across multiple users (Marshall et al., 2012). Other augmentations have aimed to make the complexity of social file management more intelligible to users, for example by allowing them to review the permissions of all shared files (S. Voida, Edwards, Newman, Grinter, & Ducheneaut, 2006), storing the history of shared files (Whalen, Toms, & Blustein, 2008), and visualising the history and permissions metadata (Rode et al., 2006); these augmentations therefore help to clarify the consequences of users' actions on other users' interactions and on the security of their own digital possessions.

In addition to aiding users in understanding shared files, increased file metadata has been used to try to improve both search and navigation. So far, this has been done both manually, by allowing users to input text annotations and images to accompany their folders and imbue them with additional meaning (Jones, Hou, Sethanandha, Bi, & Gemmell, 2010; Jones, Thorsteinson, et al., 2016), and automatically, by enriching folders with content taken from a relevant Wiki (S. Voida & Greenberg, 2009). Finally, small but ubiquitous FM actions have not been overlooked, as augmentations have aimed to: make filing new files easier by suggesting locations (Prinz & Zaman, 2005; Sinha & Basu, 2012b), improve file copying tasks by adding a many-to-one feature (Sinha & Basu, 2012c), facilitate planned backups (Cox, Murray, & Noble, 2002), and allowing multiple selection of files across simultaneously open folders (Sinha & Basu, 2012a).

Alternative FM interaction paradigms

Files and folders obviously do not exist in the computer as literal, physical paper files and folders, but are presented in this way metaphorically to provide users with a familiar idea of what digital objects are like and what can be done with them. This metaphor and the hierarchy provided with it are not the only possible way to represent and enable interaction with digital objects, and many systems have been developed to test alternative

approaches.

One theme among these systems is utilising metaphors that rely on common phenomena in human experience, such as space and time. This is achieved, for example, by putting the files into a three-dimensional space where users can arrange and automatically re-arrange (Agarawala & Balakrishnan, 2006) their documents into piles (Mander, Salomon, & Wong, 1992) and other arrangements (G. Robertson et al., 1998) in the same way they may be in physical space. This utilises the spatial metaphor already popular in modern computing while avoiding the folder hierarchy, and enables highly personalised user-made reminding cues (Bondarenko & Janssen, 2005).

Research done in information visualisation on how to display hierarchies of various kinds in efficient and usable ways is also directly applicable to the display of the folder hierarchy, and in fact folder tree structures are often the specific cases used to demonstrate various general approaches (Turo & Johnson, 1992; Xu, Esteva, & Jain, 2010). Such work has typically consisted of designing a novel approach and comparing it to various baselines (Kobsa, 2004; Merčun & Žumer, 2013), and has generally focused on visualising especially large trees (Plaisant, Grosjean, & Bederson, 2002) using various two- and three-dimensional approaches. The most prototypical of these visualisations include treemaps (space-filling rectangles) (B. Johnson & Shneiderman, 1991), of which several variations exist (Stasko, Catrambone, Guzdial, & McDonald, 2000; Turo & Johnson, 1992), and animated 3d trees (G. G. Robertson, Mackinlay, & Card, 1991).

Files may also be presented chronologically, for example by allowing the user to specify a subset of documents based on some property (time or otherwise) and presenting them as a chronologically sorted, two-dimensional array (Fertig, Freeman, & Gelernter, 1996a; Freeman & Gelernter, 1996). Both spatial and chronological representations of files entail compromise: presenting time as locations in space (on the screen) mixes metaphors, while piles are unstructured containers that are functionally identical to a flat list of folders (Treglown, 2000). Novel systems also represent digital items without metaphors, however, and typically do so simply as numerically discrete items (whether called files or otherwise) in flat lists or tables sorted by their literal properties, such as name, type, size, author,

and so on (Dourish, Lamping, & Rodden, 1999; Dourish, Edwards, LaMarca, & Salisbury, 1999a). By utilising items' properties beyond name and folder location, and the fact that users remember these additional properties (Gonçalves & Jorge, 2008a), new interactions are enabled: users may retrieve from their collection by recalling an item's narrative (Gonçalves & Jorge, 2006) or following a path of associations, such as from a user-remembered event, to an email in which it is discussed, to a document that was attached to the email (Kim, Croft, Smith, & Bakalov, 2011). Classifying by property also allows users to assign items to multiple groups, rather than a single folder (Quan, Bakshi, Huynh, & Karger, 2003), thus avoiding the single classification problem of the folder hierarchy. Items' properties can then be used to present items according to a logical division of content, such as in easily understood Venn diagrams (De Chiara, Erra, & Scarano, 2003), or in robust relational databases (Marsden & Cairns, 2003). One broad possibility enabled by focusing on item properties has been to unify digital items of all types (emails, files, Web documents, etc.) and present them together, grouped by their properties (Dong & Halevy, 2005; Dumais et al., 2003); if effective in its execution, an integrated presentation of files and documents across local storage and the Web would help to alleviate issues of information fragmentation (Bergman, Beyth-Marom, & Nachmias, 2006; Capra et al., 2014) and provide a flexibility that more closely resembles the physical world (Bondarenko & Janssen, 2005).

Another approach is to utilise properties of the user, rather than properties of the digital items, and for this user activity, task, and context have been the most popular thus far. With this approach, digital items need not be categorised in the folder hierarchy, but instead can be presented in a two-dimensional space in clusters representing their relevant activity or task (Krishnan & Jones, 2005). This can be taken even further by providing computing environments and workspaces dedicated to specific work- and PIM-related activities (Jeuris, Houben, & Bardram, 2014), where only relevant programs are displayed, and suspended while changing tasks. Demarcating a single task or activity is challenging, however; approaches to this include allowing users to generate activity names and apply them to files with tags (S. Voida, Mynatt, & Edwards, 2008; S. Voida

& Mynatt, 2009), determining an activity by analysing the times when files are in use (Krishnan & Jones, 2005), and logging the instances and times of common software interactions (Chernov, Demartini, Herder, Kopycki, & Nejdl, 2008).

As discussed above, tagging has been investigated for its potential use in providing multiple classification of files, thus obviating maintaining a folder hierarchy. Several systems have implemented this, either by using tags without the folder hierarchy (Seltzer & Murphy, 2009) or in tandem with it (Albadri, Watson, & Dekeyser, 2016). The ubiquity of the tagging concept means it can be offered as an unobtrusive feature (Oleksik et al., 2009) in both local and Web-based FM systems (Hsieh, Chen, Lin, & Sun, 2008) and in document management systems (Ma & Wiedenbeck, 2009).

Most of these novel approaches have had little effect on file management beyond their initial testing. A tagging feature has been introduced to Mac's Finder application, however, where it is offered alongside the folder hierarchy. This may be the most drastic change that file management will encounter in the near future; because current operating systems deal with files, any software that aims to replace them must still provide users with some access to them (Kaptelinin, 2003), thus prolonging the habit of managing them and therefore the need for such functionality.

A table summarising the system augmentations, alternative FM metaphors, and related hierarchy visualisation studies is presented in Table 1.3.

1.3 Theory and methodology in file management research

In this section we discuss how FM research is carried out, first by noting the current theoretical underpinnings adopted, second by examining the methods used to study user behaviour, and third by examining how systems and services are compared and improved.

System augmentation	Examples
improved cloud and file sharing (6)	Jones, Thorsteinson, et al. (2016); Marshall et al. (2012); Prinz and Zaman (2005); Rode et al. (2006); S. Voida et al. (2006); Whalen et al. (2008)
improved and assisted search (6)	Chirita et al. (2006); Ghorashi and Jensen (2012); Handschuh et al. (2007); Kim and Croft (2010); Liu and Feng (2016); Manber (1994)
enriched file or folder metadata (3)	Jones et al. (2010); Jones, Thorsteinson, et al. (2016); S. Voida and Greenberg (2009)
improved navigation (3)	Fitchett et al. (2013, 2014); Vicente and Williges (1988)
file or folder de-emphasis (3)	Bergman et al. (2009); Bergman, Elyada, et al. (2014); Lee and Bederson (2003)
improved selecting, moving, copying (2)	Sinha and Basu (2012a, 2012c)
assisted filing (2)	Prinz and Zaman (2005); Sinha and Basu (2012b)
assisted backup (1)	Cox et al. (2002)
Alternative metaphor	**Examples**
according to items' properties (14)	Adrian et al. (2009); Dourish, Edwards, et al. (1999a); Dourish, Edwards, LaMarca, and Salisbury (1999b); Dourish et al. (2000); Gifford, Jouvelot, and Sheldon (1991); Gonçalves and Jorge (2006); Haller and Abecker (2010); Hardy and Schwartz (1993); Kim et al. (2011); Mosweunyane, Carr, and Gibbins (2011); Quan et al. (2003); Sajedi, Afzali, and Zabardast (2012); Salmon (2009); Sauermann et al. (2006); Schaffer and Greenberg (1993); Thai, Handschuh, and Decker (2008)
using tags (7)	Albadri et al. (2016); Adrian, Sauermann, and Roth-Berghofer (2007); Bloehdorn, Görlitz, Schenk, and Völkel (2006); Hsieh et al. (2008); Oleksik et al. (2009); Seltzer and Murphy (2009); Voit et al. (2012b)
chronologically (7)	Fertig et al. (1996a); Freeman and Gelernter (1996); Wideroos and Pekkola (2007); Gyllstrom (2009)
spatially (6)	Mander et al. (1992); G. Robertson et al. (1998); Altom, Buher, Downey, and Faiola (2004); Bauer, Fastrez, and Hollan (2005); Agarawala and Balakrishnan (2006); Sinha and Basu (2012b)
by relevant activity (5)	Hirakawa, Mizumoto, Yoshitaka, and Ichikawa (1998); Dragunov et al. (2005); Shneiderman and Plaisant (1994); S. Voida and Mynatt (2009); Jeuris et al. (2014)
logically (3)	Bowman, Dharap, Baruah, Camargo, and Potti (1994); De Chiara, Erra, and Scarano (2003); Marsden and Cairns (2003)
integrated, combining approaches (8)	Dumais et al. (2003); Dong and Halevy (2005); Krishnan and Jones (2005); Cutrell, Robbins, Dumais, and Sarin (2006); Cutrell (2006); Dittrich and Salles (2006); Gemmell, Bell, Lueder, Drucker, and Wong (2002); Nelson (2000)
Related information visualisation studies	**Examples**
visualising hierarchies	B. Johnson and Shneiderman (1991); Kobsa (2004); Merčun and Žumer (2013); Plaisant et al. (2002); G. G. Robertson et al. (1991); Stasko et al. (2000); Turo and Johnson (1992); Xu et al. (2010)

Table 1.3 – Studies exploring FM software augmentation or alternative metaphors to files in a folder hierarchy, and studies focusing on related problems in hierarchy visualisation

1.3.1 Theoretical and conceptual frameworks

There do not currently exist any explicit theories about or theoretical frameworks for understanding file management, as there has yet to be any theory development in PIM research. This reflects a trend in information science studies generally (Hjørland, 2002), and like most IS studies, FM studies are typically justified by an appeal to a real world problem that they seek to understand or alleviate. Similarly, no philosophical positions have been discussed in relation to FM or PIM, and the predominant implicit position of FM research is that of science more generally: post-positivism. Put very briefly, this position takes human perceptions and scientific measurements to be of a *real* world where causes reliably determine effects but various biases are taken seriously. This position is generally assumed when using quantitative approaches to scientific inquiry. By contrast, a constructivist position, which holds that the world is *constructed* by and consists only of perceptions and interpretations, is typical of qualitative approaches seeking to identify how meaning and behaviour are constructed and conceived of (Bryman, 2012). Both approaches may be useful in FM depending on the research questions being asked, as may the many positions in the spectrum between the two, but careful consideration of how these influence the questions, methods, and conclusions of FM research has yet to be carried out.

There are, however, two models for characterising user behaviour in PIM, and these include mention of and apply straightforwardly to FM contexts. Each characterises user behaviour as belonging to one of three categories; for Jones (2007a), these categories are keeping, finding or refinding, and organising (also called metalevel) behaviour, while for Whittaker (2011) these are keeping, exploiting, and managing. The two are obviously similar: since exploiting or utilising information often entails (re)finding it, those categories could be collapsed into one (e.g., refinding and utilising), making the approaches essentially equivalent. Alternatively, exploiting and refinding could be kept distinct and serialised (e.g., one refinds and then utilises information), making the approaches complementary. Regardless, these frameworks capture the main concerns of traditional PIM and FM, as are reflected in the themes in user behaviour as summarised above. Notably miss-

ing from each framework, however, is explicit mention of the increasingly social aspect of personal information, which consists not only of co-managing (captured by metalevel or managing) but also of sharing (i.e., sending, receiving, and so on) information.

The conceptual frameworks and as-of-yet inactive theoretical landscape of FM are summarised in Table 1.4.

Concept	Summary
Theories, philosophical positions	*There has not yet been discussion of theory or philosophical positions as they relate to file management research. Philosophical positions are generally implicit, and either post-positivist in quantitative studies or constructivist in qualitative ones.*
Author	Conceptual framework of PIM
Jones (2007b)	keeping, (re)finding, and managing (metalevel) information
Whittaker (2011)	keeping, exploiting, managing information

Table 1.4 – Conceptual and theoretical frameworks that have been discussed for PIM and are applicable to FM.

1.3.2 Methods for understanding user behaviour

Three general approaches to studying FM behaviour can be identified in the literature, and are often used together: ask participants about their behaviour, observe the behaviour directly, and infer the behaviour from the file system. We examine each in turn.

Asking

Asking participants about their file management behaviour has typically been done to discover user behaviour and challenges and understand the relevant contexts, usually by capturing participants' responses with digital questionnaires or recorded interviews. For example, studies using this approach have examined the challenge of coordinating files across multiple devices (Capra, 2009), difficulties in managing files in Mac OS (Ravasio et al., 2004), students' habits in downloading documents (Huvila et al., 2014), opinions about graphical file management (Kaptelinin, 1996), and user perceptions about searching for files (Teevan et al., 2004; Bergman, Beyth-Marom, Nachmias, Gradovitch, & Whittaker, 2008). It is rarely the only approach used in a study; rather, it is combined with the approaches described below when user perceptions are needed to understand the

observed or inferred behaviour (Whitham and Cruickshank (2017) for example, combine all three approaches).

This approach is direct, as data about user perceptions and behaviour can be gleaned from participants rather than inferred from their behaviour. As with other forms of ethnographic study the data collected can be rich and useful for understanding contextual factors and informing the design of relevant systems and services. One disadvantage, however, is that users may not have accurate knowledge about their own behaviour: one study found a large discrepancy between users' attitudes about tagging their files (e.g., very positive) and their actual tagging behaviour (e.g., they typically did not tag files even when a good tagging system was presented and explained to them) (Bergman, Gradovitch, et al., 2013b). Further, they may simply not be aware of any number of details about their own behaviour; for example, it is unlikely that anyone is cognisant of the number of redundant files they keep.

Observing

Observation is a popular approach to investigating fine-grained phenomena and specific challenges in FM. Studies using this approach have, for example, sought to understand if digital documents are organised like paper documents (Barreau, 1995), how information from the Web is stored in files (Jones et al., 2001, 2002), and various challenges of file retrieval (Bergman, Whittaker, & Falk, 2014). This is typically done by recording participants as they perform structured tasks, ordinary work, or a guided tour, where they navigate and explain their folder arrangement to an observer and perform common file management tasks along the way. Observation notes stored on paper, video recordings, and screen shots are all relatively simple methods that have been used to capture data, although unclear recordings have resulted in lost data (Bergman et al., 2010). Complex methods for observing users in more fine-grained ways have also recently been used (Benn et al., 2015).

Using this approach, actions are observed as they occur semi-naturally (e.g., during work) or when solicited (e.g., in a structured task or guided tour). Observation always

takes place during some time, however, and thus necessarily does not see what participants are doing when not observed. This may be alleviated by supplementing observations with logs and inferences drawn as discussed next; for example, file creation, access, and modify times stored by the operating system can provide evidence of what participants do between immediate observations.

Inferring

Users' actions determine properties of their file systems and the files and folders; for example, the folder hierarchy depth, the types of files stored, and the size of the collection in bytes and in total files and folders are each the result of specific user actions to store and organise their digital items, and their properties provide traces of this behaviour. The file system therefore serves as an artefact from which we may infer users' past behaviour, and studies have used this to study FM since the 1980's. They have, for example, observed files' sizes (A. J. Smith, 1981) and names (Carroll, 1982), examined how files are organised into folders (Khoo et al., 2007), explored the role of provenance in file retrieval (Jensen et al., 2010), studied the document management behaviour of students (Henderson & Srinivasan, 2011), and examined the effect of folder depth on file retrieval (Bergman et al., 2012).

This approach has been implemented in two ways, which are used roughly as frequently and sometimes together. First, researchers have examined the file system as it appeared in the recordings of the interviews, guided tours, or structured tasks described above. This method is relatively simple and does not require technical skills like programming, but given the large number of observable file system properties discussed below, manual notation of the properties is necessarily either highly laborious to collect and analyse or else limited to a small number, and it may does not capture properties of portions of the file system not seen during the task or tour (Bergman et al., 2010). A second way is to use custom software to traverse the folder tree, recording data about the files and folders encountered, or to log user actions or changes to files and folders.

Automated methods facilitate studying a large sample and many variables (e.g., file

system properties), including temporal data, but are a technical challenge to develop and implement (Dinneen, Odoni, Frissen, & Julien, 2016). Both manual and automatic collection methods require participant trust to let the researcher, possibly perceived as an expert in PIM, see their digital organisation or perceived lack thereof (Barreau, 1995). Automatic methods may also require researcher supervision to use, thus restricting sample size by being difficult to administer, or may be an obstacle to recruitment because it is difficult to find users willing to expose and share their digital possessions and desktops, entailing that participants are from an available but niche group like trusting colleagues. It is also difficult to develop such software to support multiple operating systems; perhaps as a result, researchers have instead relied on tools packaged with the OS (as in, for example, Evans & Kuenning, 2002) that provide minimal functionality, and typically focused on a single OS (as in, for example, Khoo et al., 2007).

A look at thirty-one studies examining the file system reveals the use of these methods, the number of participants in the sample, and the file system properties examined (presented together in Table 1.5). It should be noted that a low number of file system properties or a small sample size does not necessarily indicate an ineffective FM study or researcher oversight, as studies have explored differing research questions requiring collecting data about only particular file system properties.

Twenty-eight properties of the file system have been examined across the studies mentioned above, regardless of the data collection method used. This includes five variables that are particularly important to general PIM contexts: collection size, folder depth, folder breadth, folder size, and redundancy (e.g., in file and folder names) (Bergman, 2013). Twelve additional properties were suggested by Dinneen, Odoni, Frissen, and Julien (2016), resulting in forty properties available for use in FM research (presented in Table 1.6).

Together, these properties characterise each category of FM behaviour discussed above, and in smaller groups provide insight into particular actions and challenges users regularly encounter. The most commonly made measurements include folder tree height, breadth, number of subfolders per folder (sometimes called *branching factor*), and con-

Study	n =	Data collection methods	FM properties examined
Satyanarayanan (1981)	8	simple software	collection size; file size
Carroll (1982)	22	structured task	file type; collection size; file name; length of name
Akin, Baykan, and Rao (1987)	171	structured task	branching factor; folder fullness; folder depth; file and folder names
Bennett, Bauer, and Kinchlea (1991)	3	simple software	collection size; file size; use of symbolic links; file types; number of folders
Sienknecht et al. (1994)	267	simple software	file size; collection size; files per folder; branching factor; file access
Barreau (1995)	7	guided tour	file names; file access times; use of default locations
Nardi et al. (1995)	15	guided tour	file type (ephemeral, working, or archive)
Douceur and Bolosky (1999)	10,568	simple software	file size; files per folder; folder depth; file creation and modification; file types; leaf folders
Vogels (1999)	45	simple software	file size; file type; collection size
Downey (2001)	562	simple software	file size
Evans and Kuenning (2002)	22	simple software	file type; file size
Gonçalves and Jorge (2003b)	11	simple software, interview	tree depth; total file count; branching factor; files per folder; file types; file size; file creation, modified, accessed times; use of numbers, whitespace, and punctuation in names; length of file names; use of shortcuts/symlinks
Boardman and Sasse (2004)	31	simple software, guided tour, diary	total folders; folder depth; unfiled files
Ravasio et al. (2004)	16	guided tour	file age; files per folder; use of desktop
Henderson (2005)	6	simple software, interview	total folders; file names; duplicate file names; duplicate folder names; branch consistency
Jones et al. (2005)	14	guided tour	branching factor; file types; file names
Agrawal et al. (2007)	62,744	simple software	file size; collection size; file types; file creation and modification; files per folder; use of default locations; file depth; folder count
Khoo et al. (2007)	12	simple software, interview	use of default folders; roots per user; use of desktop; tree height and breadth; files per folder; file names
Hicks et al. (2008)	40	simple software, questionnaire	file names; tree depth; file depth; file size; collection size in bytes; file types; file and folder duplication (by name); file access times
Hardof-Jaffe et al. (2009b)	518	custom online environment	collection size; tree dimensions; files per folder; file depth; unfiled files
Henderson and Srinivasan (2009)	73	simple software	collection size; tree height; file depth; branching factors; root folders; file name duplication, folder name duplication; empty folders
Bergman et al. (2010)	296	structured task	file depth; use of desktop; use of shortcuts; files per folder; branching factor; use of default locations; files per folder
Henderson (2011)	73	interview	unfiled files; tree height; file name duplication; folder name duplication; use of desktop; use of default locations
Henderson and Srinivasan (2011)	10	interview, simple software	unfiled files; tree height; file name duplication; folder name duplication; use of desktop; use of default locations
Bergman et al. (2012)	289	structured task	file depth; files per folder; branching factor
Bergman, Whittaker, and Falk (2014)	275	structured task	file depth; file type; file access time
Massey et al. (2014)	62	simple software	total files; use of desktop; file types
Zhang and Hu (2014)	12	guided tour, simple software	tree breadth, tree shape, files per folder, branching factor, total files, folder depth
Fitchett and Cockburn (2015)	26	interview, logging	file access; file types; file name length; file depth; use of desktop
Benn et al. (2015)	17	structured task	folder depth
Whitham and Cruickshank (2017)	12	guided tour, scan and logging software	total files, total folders; file and folder access and modify times;

Table 1.5 – Studies observing participants' file systems, number of participants, data collection method, and file system measures reported.

sistency (usually defined as deviation of branches from the average), which inform us of, respectively, the maximum depth to which users may need to traverse to find a file, the maximum and average number of navigation decisions at any depth they may need to make at any depth, and the likelihood of the user encountering an unfamiliarly structured area (or branch) of the tree during navigation. The *time of last access* of files and their depth in the folder hierarchy can help to quantitatively describe users' archiving habits, and the number of duplicated file and folder names indicate the difficulty they face in differentiating and naming similar items in their collections (Henderson, 2011). Properties can also be examined for correlation with individual difference and external factors, for example to see if certain occupations or personality styles correlate with the average length of file names or total number of files (Massey et al., 2014). The varied goals and research questions present across studies of this type entail that despite collectively looking at many of these properties, a complete quantitative description of general FM behaviour (i.e., storage, organisation, retrieval...) does not yet exist and cannot be derived from cross-study analyses (e.g., meta-analyses). The implications of and suggested solution to this are discussed in the future research directions, below.

FM topic	Data about	Properties
Storage	Hardware (4)	# of available drives, hard drive capacity, use, and free space;
	Collection (13)	total files, total folders; collection size (in bytes), collection size (files + folders); file extensions/types; file sizes; file age, folder age; shortcuts/symlinks, hidden files, hidden folders; duplicate files (by hard link), duplicate folders (by hard link)
	Semantics (7)	File or folder name, length of name, numbers in names, punctuation or special characters in names, duplication of names; Letters in names, whitespace in names
organisation	Structure (12)	Root folders; tree breadth, tree depth; folders in each folder (branching factor), files in each folder; file depths, folder depths; branch consistency or skewness; use of desktop for storage, use of default folders; inaccessible folders in user space; folders excluded by participants from study
Retrieval	File access (4)	File access times, file modify times; folder access times, folder modify times

Table 1.6 – 40 properties of file system, measured to infer participants' FM behaviour

1.3.3 Methods for designing and evaluating FM systems

As noted above, the general process for improving existing and novel FM systems and approaches entails evaluating systems' performance or users' performance or preference, for example during structured tasks and in comparison to some baseline system. However,

it is agreed among researchers that meaningfully evaluating and comparing PIM systems is extremely challenging (Kelly, 2006), due to four factors that apply as much to FM as they do to broader PIM contexts. First, PIM behaviour is complex and idiosyncratic, so the relative effects of the many factors can be difficult to understand and it is not always clear which tasks are best for an experiment (Capra & Perez-Quinones, 2006). This is compounded when performing longitudinal studies, as user behaviour across time is not well understood; longitudinal approaches to evaluating FM are thus rare (Dinneen, Odoni, & Julien, 2016). Second, representative data sets do not exist; a representative file and folder collection has not been identified, and so a standardised data set for testing does not yet exist, nor does a model of the average relevant actions and activities. It may soon be possible to create representative collections by extensively recording file system properties across many participants (Dinneen, Odoni, Frissen, & Julien, 2016), and to create representative models of user behaviour by combining file system properties with activity logging (Chernov et al., 2008). Third, traditional evaluation measures do not apply straightforwardly to FM contexts; for example, recall and precision are of limited use in FM retrieval evaluation, as most FM retrievals are looking for a particular file rather than a large batch of files (Fitchett & Cockburn, 2015), and it is impractical to ask a single participant to make relevancy judgements for all of their documents and invalid to ask third parties to help in this (Gonçalves & Jorge, 2008b). Fourth, though it is essential for carrying out valid comparative evaluations, it can be difficult to make fair comparisons between systems and approaches when they are created with differing affordances and intended interactions (Voit, Andrews, & Slany, 2012a). One approach to comparing efficacy, efficiency, and usability across disparate systems is by doing an evaluation called GOMS model analysis (Kieras, 1999), which can provide an outcome-based comparison in cases where possible user behaviour can be enumerated and predicted with some confidence. This has been used, for example, for testing the efficiency in moving and deleting files in a new file manager as compared with the existing File Explorer (Sinha & Basu, 2012a).

Evaluation aside, an explicit approach to the general design of PIM systems that

clearly applies to FM systems is the *user-subjective approach* (Bergman, Beyth-Marom, & Nachmias, 2003), which emphasises that PIM tool design should be concerned with what the users find important, rather than studying only how users behave with current, limited systems. This approach has been explicitly utilised in several studies (Bergman, Beyth-Marom, & Nachmias, 2008; Bergman et al., 2009; Bergman, 2012), and so shows promise for FM-specific software design. Examples of literature pertaining to reflection on the design and evaluation of FM systems are presented in Table 1.7.

Topic	Examples
System design	Bergman et al. (2003); Bergman, Beyth-Marom, and Nachmias (2008); Bergman (2012)
Experiment, task, data set design	Capra and Perez-Quinones (2006); Chernov et al. (2008); Dinneen, Odoni, Frissen, and Julien (2016); Dinneen, Odoni, and Julien (2016); Gonçalves and Jorge (2008b); Kelly (2006); Voit et al. (2012a)

Table 1.7 – Examples of literature relevant to the design and evaluation of FM systems

1.4 Discussion

We discuss here the importance and relation of FM research to various other research areas, and then discuss the future directions and challenges facing FM research.

1.4.1 Importance to other research areas

By virtue of studying how humans use computers to manage information, FM research shares the concerns and methods of research areas such as personal information management, computer supported collaborative work, information retrieval, and human-computer interaction. It also has broad import for core subfields in information sciences like information behaviour and organisation, and personal archiving. Finally, it even has overlap and potential implications for psychology, computer science, and philosophy. We discuss these in turn.

The research area most closely related to FM research is personal information management (PIM), which can be argued to be the broader *parent* topic to which FM research belongs, although this has not yet been explicitly posited and defended. For example, in

this view FM can be seen as a subset of PIM focusing specifically on how people manage information at the file and folder level. The contexts of files and folders is arguably of crucial importance in PIM, given that much of the information of our daily lives resides in the digital domain, specifically in files. Indeed, the categories of research described above could be used to describe common concerns in PIM: to understand peoples' behaviour when personally managing information, to understand what gives rise to differences in this behaviour, and to improve the design of the relevant systems and services that support this. Typical FM activity accords with the various conceptions of *personal* fundamental to and used in PIM literature; for example, that *personal* includes being controlled by, owned by, about, directed toward, sent by, experienced by, or potentially relevant to an individual (Jones et al., 2015). Perhaps unsurprisingly then, files and folders have been present in and are relevant to many PIM studies that focus on the management of digital items by type or format, including digital music collections (Brinegar & Capra, 2010, 2011), digital photo collections (Rodden & Wood, 2003), and scholarly references (Fastrez & Jacques, 2015). Though users have the option to manage these digital items within their respective format-specific applications, they may also manage them as files, and insights gleaned from FM studies have implications for their general management. More about these two modes of management, of digital items as files or as specific formats, is discussed below.

FM research has relevance to human-computer interaction (HCI), information retrieval (IR), and computer-supported cooperative (or collaborative) work (CSCW), and this is reflected in the presence of PIM workshops in the last decade at the relevant HCI (2008 at SIGCHI, 2016 at CHI), IR (SIGIR 2006), and CSCW (2012) workshops. Managing files is a required activity for anything beyond the simplest computer usage. Due to this ubiquity and fundamentality it is of considerable relevance to research in HCI, where the file-folder metaphor has been a common example of typical user interactions, for example in debates about digital design and manipulation philosophies (Frohlich, 1997) and the broader desktop metaphor (J. Johnson, 1987; Ravasio & Tscherter, 2007). It is within the HCI community primarily that the debate about the use of the file and

folder metaphor, summarised above, has taken place. FM may also serve as an excellent context for advancing our knowledge of information foraging theory (Pirolli, 2007), which is of interest to those studying HCI and information behaviour (IB) alike; with folders and files serving as metaphorical bushes and berries, it is reasonable to describe users' FM behaviour as enriching their file systems by storing and organising, following scents by navigating, and foraging by retrieving.

Because much FM activity consists of retrieval or is done to support later retrieval, it is perhaps unsurprising that FM research also has a close connection with IR research. The role of search (both for files and through files) in FM has been a focal point of FM research, and this has provided insights into how users retrieve files with search, navigation, or both, as discussed above. FM systems and their users benefit from innovations in IR research, for example in the retrieval and ranking algorithms and improved full-text and faceted search. FM is also relevant to research into CSCW and a topic within PIM known as group information management (GIM), as the opportunities, challenges, and implications of co-managing shared files, especially for collaborative work, are likely generalisable to broader contexts. For example, a study (Bergman, Whittaker, & Falk, 2014) of the impact of shared files on retrieval success participates in and has implications for FM in understanding users' refinding behaviour, IR in supporting user behaviour with better file search algorithms, and CSCW in understanding how the shared files have supported shared tasks.

FM research also has relevance to core areas in information studies, such as IB and information-seeking behaviour (ISB), as is reflected by the presence of two PIM workshops at the ASIS&T annual meeting (in 2009 and 2013). IB research, understood as investigating "how people need, seek, manage, give, and use information" (Fisher, Erdelez, & McKechnie, 2005, p. xix), is clearly related to both PIM and FM, where users create, manage, and retrieve information stored in files, thus exhibiting particular patterns of IB. Thus unsurprisingly, typical IB patterns like filing, archiving, and organising collected information (Meho & Tibbo, 2003) match very closely what users do with files as described in the FM strategies previously characterised (Berlin et al., 1993; Malone,

1983; Boardman & Sasse, 2004). The role files play in greater IB and ISB patterns has been touched upon tangentially in many studies of PIM and ISB, but given the prevalence of files this should be investigated further; changes in ubiquitous and fundamental information software such as a file manager will likely affect the information behaviour of various groups.

File management research also has a clear but so far largely implicit overlap with work in personal archiving (PA) or personal digital archiving (PDA), which will become clearer after the Personal Digital Archiving 2017 conference to be held at Stanford University, wherein PIM will be discussed. The extent of the importance of the two fields for each other is implicit in the following description of the concerns of PA activity and research: "what we have written, what we have read, where we have been, who has met with us, who has communicated with us, what we have purchased, and much else [that] is recorded digitally in increasingly greater detail in personal digital archives, whether they are held by individuals, institutions, or commercial organisations, and whether we are aware of those archives or not" (Hawkins, 2013, p.2). For those digital archives that are personal in virtue of being managed or owned by some person, it is very likely that FM is taking place, and is either being done neglectfully, thus under-facilitating later reuse, or painstakingly, and could thus benefit from thoughtfully designed software support. Indeed, numerous studies consider a person's files as being part of their personal archive or digital possessions collection (Kaye et al., 2006; Marshall, Bly, & Brun-Cottan, 2006; Siddiqui & Turley, 2006; Marshall, McCown, & Nelson, 2007; Cushing, 2013; Massey et al., 2014). That some files are regarded differently than others and are kept and preserved across a long span of time is certainly of interest to FM research, and it is clear that, say, file management augmentations could be designed specifically to support personal archiving. It is therefore unsurprising that the potential for PIM studies, including how people manage files, to be used to better understand personal digital archiving has already been suggested (Bass, 2013). The above quote also demonstrates the distinction between PA and FM concerns, however: FM is necessarily not concerned with physical objects, and traditionally has not been concerned with digital collections that are about a person

but not managed or owned by them.

FM research also has potential import for research in knowledge organisation (KO), which is concerned with the nature and quality of knowledge organisation systems used to organise documents. Labelled folders and their parent-child relationships present users with a free-form way to structure and name information as they want to, and so identifying how and why they do so may produce insights for KO systems design in general. Identifying trends across adequate numbers of users would mean establishing reflections of current practices and expectations of document organisation tools (folder trees in this specific case), which should be considered when designing KO systems. For example, knowing what is the mean depth and breadth of a group of users' folder trees suggests the shape of KO hierarchies users are accustomed to browse and navigate, which is an open question for KO structure interface design (Julien, Tirilly, Dinneen, & Guastavino, 2013). At the theoretical level, research in KO has been concerned with, among other things, finding confirmations of power law distributions (Smiraglia, 2002), and these are likely to be found in file systems as well, where most folders would contain small numbers of files while a small group of folders would contain most files. The field of KO has only begun to focus on individual differences (Rowley & Hartley, 2008), and so this may be a valuable research direction that FM research could aid: what factors determine which information structures a user is more comfortable with, and is the phenomenon similar in the FM context?

PIM (and FM) behaviour is also of explicit concern to those looking to improve library services (Fourie, 2011, 2011; Otopah & Dadzie, 2013). Between this concern and the relevance FM has for the fields discussed above, particularly personal digital archiving and information behaviour, FM is therefore of concern and broad import in IS.

Moving only slightly further afield, we think it is reasonable to infer a possible relevance of FM to computer science, where a considerable body of existing literature aims to understand the contents and access patterns of file systems, such as file size distribution (Tanenbaum, Herder, & Bos, 2006), to optimise hardware, firmware, and software. FM studies focusing on real-world file systems that users have interacted with may provide

valuable data sets for such design goals, especially given that most of such computer science studies have examined atypical contexts like servers.

But FM research and the file-folder paradigm may also be useful in fields beyond those concerned with the information and information systems. We have discussed above the psychological aspects of FM previously examined, but the relevance of FM to psychology may extend beyond this. For example, metaphors, metaphorical thinking, and categories and categorical thinking are common objects of study in psychology and prominent in FM (digital and analogue) and PIM generally (Case, 1991). It is not a stretch to think that other dimensions of individual difference are factors in FM, including those concerning psychology, like cognitive styles, and decision making processes (Kozhevnikov, 2007). At its broadest, general trends in file management studies may also be of interest to those studying topics like Philosophy of Information and Philosophy of Computing, which seek to understand what is possible in the digital realm, how much information we are storing as a society (Lyman et al., 2003), and to what extent humanity has moved into the infosphere (Floridi, 2010).

FM therefore has significance and potential import for many fields, including several within information science and surrounding HCI. Table 1.8 lists the fields and disciplines, discussed in this section, that have connections to FM research. We next discuss the future of FM and its study.

Field or discipline	Abbreviation
Computer science	CS
Computer-supported cooperative work	CSCW
Group information management	GIM
Human-computer interaction	HCI
Information behaviour, information-seeking behaviour	IB, ISB
Information retrieval	IR
Information science or studies	IS
Organisation of information, knowledge organisation	OI/IO, KO
Personal archiving, personal digital archiving	PA, PDA
Personal information management	PIM

Table 1.8 – List of fields and disciplines connected to file management research, with field abbreviations used in this paper.

1.4.2 Future challenges and research directions

In this section we present a discussion of the future challenges and directions in FM research that is structured to reflect the existing areas of research identified above, and have included at the end a discussion of the future of files and their management systems.

Improved understanding of user behaviour

Future research into users' behaviour will likely benefit from combining complementary insights from qualitative and quantitative studies, providing a complete picture of the various aspects, scope, and contexts of behaviour. For this, a broad quantitative description of typical behaviour (i.e., of many measures of FM behaviour) could complement the rich characterisations of users' FM behaviour that has emerged from the many qualitative descriptions and unify the disparate quantitative descriptions discussed above. This would enable advanced methods for understanding such behaviour, like principal component analysis, user modelling, and the generation of a standardised, representative data set for FM system evaluation (Chernov et al., 2008).

Deriving such a quantitative picture from the findings of previous studies is currently impossible, however. As noted above, one consequence of the varying goals and research questions of previous studies is that many study contexts are fundamentally incomparable; for example, where one study examines the retrieval of recently used files seen during a controlled experiment (Bergman et al., 2010), another examines the folder structures created by students in a proprietary, online environment during a class assignment (Hardof-Jaffe, Hershkovitz, Abu-Kishk, Bergman, & Nachmias, 2009a), and it is difficult to compare results across studies of public computers (Vogels, 1999) and servers (Sienknecht et al., 1994), or of only media files (Evans & Kuenning, 2002), shared files (Bergman, Whittaker, & Falk, 2014), or recently accessed files (Fitchett & Cockburn, 2015). Another natural consequence of studies' varying goals is that even when contexts have been comparable, the measures collected and reported have typically differed; for example, studies of the file system have collectively looked at 28 of 40 or more potential file system measures (Dinneen, Odoni, Frissen, & Julien, 2016), but typically with few

measures per study (mean 4.4), and whereas one study reports (among other measures) the maximum depth at which folders are stored (Henderson & Srinivasan, 2011), another reports the average depth of currently used files (Fitchett & Cockburn, 2015).

While the quantitative description outlined here cannot be derived from existing studies, it could be the explicit goal of future studies; specifically, future studies may examine as many of the available file system measures as possible and in as many contexts or as general of a context as possible. Such studies will require robust and capable data collection tools that overcome the limitations of current tools identified above. Once a more complete understanding of FM behaviour is achieved, fuelled by both qualitative and quantitative insights, it may be useful to investigate how user behaviour differs in similar contexts, like the management of Web browser bookmarks (Kaye et al., 2006) or emails (Ducheneaut & Bellotti, 2001; Kalman & Ravid, 2015; Mackenzie, 2000).

Time remains a challenge for understanding FM behaviour. Some of what is known about FM behaviour was established in studies that are now dated and possibly obsolete; for example, several (10) of the file system studies described above took place twenty years ago or more when graphical interfaces were relatively new, storage was expensive, and file name limits were much shorter. Though the essential nature of file management has not changed, several aspects of it have (e.g.: cloud storage), and this is reflected in the older studies, which, for example, aim to determine the optimal moment to archive working files on tape (A. J. Smith, 1981). Looking forward, although long-term management is a general concern of PIM (Jones, Bellotti, et al., 2016) few FM studies have been longitudinal, and implementing such studies is difficult (Dinneen, Odoni, & Julien, 2016).

Improved understanding of determinant factors

From the above summary of research into the individual differences and external factors influencing FM behaviour one may reasonably conclude that further research is needed to understand and support for these factors. Factors like occupational traits, task and information type, the operating system, computer literacy, spatial ability, and personality style are not yet well understood, but may play significant roles in how users struggle

or succeed in managing their files. Even the principled differences between the OSes in how users *can* manage files has not been made explicit. The default FM presentation style differs between the OSes, and this seems to affect the retrieval of recently used files (Bergman et al., 2010), but what of other system-based differences? Most of the details of these differences are scattered across user manuals and release notes, and have not been at the forefront of FM research despite their obvious influence.

The effects of individual differences on FM behaviour are also good candidates for future FM research, as no specific difference is well understood. For example, the two previous studies of spatial ability suggest that file management is influenced by general spatial cognition (Vicente et al., 1987; Vicente & Williges, 1988), but it is unclear if this extends beyond folder navigation (e.g., to folder organisation) and to what extent spatial ability specifically is responsible for such influence. The relationship between personality and file management also remains unclear, as discussed above, and additional individual differences are perhaps even more likely playing determining roles in FM. For example, one promising difference to investigate is cognitive style, the general way people think about information (Sternberg, 2008), which has been studied for how it affects learning (Tsianos, Germanakos, Lekkas, Mourlas, & Samaras, 2009), decision making (Kozhevnikov, 2007), information seeking behaviour (Ford, Wilson, Foster, Ellis, & Spink, 2002), Web browsing (Chen & Rada, 1996), and Web search behaviour (Hariri, Asadi, & Mansourian, 2014; Kinley, Tjondronegoro, Partridge, & Edwards, 2014). It is reasonable to infer that FM behaviour may be influenced by cognitive style, and in particular Riding's view of cognitive style, which integrates several views (R. Riding & Cheema, 1991), may be useful for examining this; it defines cognitive style as a preference for verbal- or image-based and analytic or wholistic information and thinking (R. J. Riding, 1997). This seems well-suited to studying FM, where users have opportunities to act on these styles and producing file and folder arrangements that reflect their style, for example by categorising files with many folders or synthesising them into a few, or by relying on folder names or images for retrieval.

Improved systems and services

Applying the findings of previous studies to improved systems is a fertile area for future FM research. One direction for this is in helping users understand, whether analytically through information literacy or intuitively through system transparency, the FM metaphor and FM system capabilities. Users often do not understand files, digital content, the actions that can be done with a file, when those actions are appropriate or reliable, or who owns and can access a file (Brostoff et al., 2005; Odom, Zimmerman, & Forlizzi, 2011; Harper et al., 2013). Future systems should therefore not only be faster, enabling greater productivity, but also simpler, either enabling more accurate and easier mental modeling or precluding the need for it. This, in turn, requires identifying specific confusions.

The development of future FM systems may be guided by existing considerations and opinions, for example that systems should improve existing systems by facilitating flexible *ad hoc* restructuring (Bondarenko & Janssen, 2005), and act as a prosthesis for human memory and support intuitive and natural interaction (Trullemans & Signer, 2014b). The usabilty of FM software has not been previously touched upon, and is thus a promising direction for improving FM systems. This may be achieved, for example using a GOMS model (Goals, Operators, Methods, Selection rules) or by designing FM software for specific uses or user groups, such as new and casual computer users (Sinha & Basu, 2012a).

Though it has been identified as mainly supporting navigation rather than replacing it, search will likely continue to be an important research area. There are many potential improvements to be made to search, for example by improving the display and interactivity of file search results (G. Smith et al., 2006), further integrating search with navigation by using queries to guide navigation (Fitchett et al., 2014), or further still, creating two-way interactions between the file tree in and search results as has been done with LCSH by (Julien et al., 2016). The evaluation of desktop search, where recall and precision are imperfect measures for reasons discussed above, may find benefit in the application of alternative measures, like mean reciprocal rank.

Building systems to support GIM and the social aspects of FM is promising. Currently, Dropbox and such software allows for synchronisation of individual file spaces, but as discussed above users often misunderstand where exactly these files are and what can be done with them. Something like the Dogear social bookmarking system (Millen, Yang, Whittaker, & Feinberg, 2007), but with successful integration of files, would likely be valuable in supporting users in tasks requiring collaborative FM. Views, or on-the-fly, ephemeral display of sets of folders and files, may help with this and with overcoming problems of mutual intelligibility (Dourish, Lamping, & Rodden, 1999), especially if unfamiliar folder structures are modified with hierchy pruning algorithms (Julien et al., 2013), but it remains unclear if such systems would do more to enable or confuse users.

Designing and improving services, such as library services (Fourie, 2011), to support PIM and FM is a promising but difficult future research direction. Though information literacy and education initiatives may be designed to include FM and other aspects of PIM, it is first necessary to identify best practices so that recommendations can be made. That few prescriptions are derived from PIM research is surprising given the vast array of strategies for categorising and filing paper records and documents that were present and promoted in the 1980's (Gill, 1988) and the subsequent proliferation of the digital computer file; some individuals now have more files in the digital domain than organisations had on paper, but fewer resources for organising them. These office file management strategies of the past may serve as inspiration for future digital organisation strategies, as they could accommodate a wide range of organisational approaches, but would need to be updated to account for the current, digital format, and subsequently tested comparatively to establish their relative efficacy.

To our knowledge, no studies have examined FM in explicitly mobile contexts. Such research and the development and testing mobile FM systems therefore remain open and important directions for future work. For many users, cell phones and other devices exist at a liminal but real space between casual media consumption and intensive computing where important emails are received containing file attachments to be downloaded, edited, backed up, and uploaded. But support for this is currently minimal, with the existence

of files being somewhat of a secret. For example, one can, in fact, access the files on an Android device by downloading a file manager; doing so may reveal previous downloads from the Web that have been saved into a folder called Downloads. These can then be renamed, moved, organised, edited, attached to emails, deleted, and so on. It is currently unknown to what extent mobile device users are doing FM, but given the different interaction affordances of mobile devices, the activity may entail a unique set of challenges (Tungare & Perez-Quinones, 2008). Alternative FM contexts, such as mobile FM, should therefore be investigated and designed for, and this will become especially important as the Internet of Things comes into fruition; for example, if your smart fridge takes a photograph, is the file stored on the fridge? In a folder? Will it later be moved a to another device, and then into a folder? Can it be copied or renamed? How will users conceive of the photo if it is not presented as a file? It is a reasonable concern that the less a device resembles a traditional computer, the more complex it will become to understand and perform FM with that device.

Improving theory, concepts, methods

FM concepts, models, and theories all stand to benefit from refinement in future research. Even the most basic concepts used in FM research can and have reasonably been debated for their precise definitions and general usefulness and vocabulary (Harper et al., 2013): what is a file, what can be done with it, and how should we talk about it? This is no trivial task, as understanding and defining digital objects is incredibly challenging (Hui, 2012), but may be essential if we hope to present clear concepts to FM system users.

The advancement of useful models and theories is obviously a desirable direction for FM and PIM research. One approach may be to adapt existing theories, such as information foraging theory (discussed above) or the records continuum model (Huvila et al., 2014), while another is to generate theories from the data. So far, both approaches remain relatively untouched.

We have noted above that most FM studies implicitly employ post-positivist or constructivist epistemologies. Though we do not see obvious problems this might imply for

particular studies, differing philosophical assumptions do lend themselves to different interpretations of findings (Hjørland & Hjørland, 2005), and so FM researchers would likely benefit from awareness of this. An explication of how different epistemologies would result in different findings in FM research specifically may not be essential at this time, but it would certainly be enlightening. Perhaps more immediately needed is a careful articulation of the relevant ethical concerns and positions about data collection and management in PIM research and the design of PIM software (Ferguson, 2016).

We discussed above the nature and limitations of various approaches to collecting data about users' FM behaviour (i.e., asking, observing, inferring). Dedicated efforts to improve data collection tools may help to overcome such limitations to the benefit of future research. For example, we noted above that the tools currently available for collecting quantitative data (i.e., used to infer user behaviour) do not collect data about many of the available file system properties, and that they are difficult to administer. Should new tools or improvements to existing tools be developed, sharing these for reuse in FM research would benefit the field.

Recently, more advanced technology (fMRI) has been adopted (Benn et al., 2015) for observing users, and software that facilitates collecting data about the file system has been developed and shared (Dinneen, Odoni, Frissen, & Julien, 2016). Future research may further benefit from considering adapting sophisticated methods from HCI and computer science research, like logging and system traces, to record fine-grained data about user behaviour that a file scan does not reveal, like file open times and changes in the size of particular files (Ousterhout et al., 1985; Baker, Hartman, Kupfer, Shirriff, & Ousterhout, 1991; Roselli, Lorch, & Anderson, 2000).

The future of files and FM systems

In time, the ideas tested in FM prototypes trickle into both commonly used software and specialised PIM software (Kljun et al., 2015b). This fact and the research areas described above might together imply that over the coming years FM software will simply continue to improve incrementally until all FM is performed optimally. But these improvements

have come slowly and require a more detailed description of FM behaviour and its component and determining factors than is currently available, and changes in computing initiated by software developers may well modify or replace the FM metaphor before such knowledge is identified. Preliminary conceptions and rumours about such ideas have lead to some common questions (e.g., at scholarly conferences) about the future of file management, including:

- Desktop search is improving and my Mac now comes with support for tagging; won't this solve all of our FM problems and preclude the need for folders?
- I don't organise my music because iTunes does it for me; can't we take the same approach with every file format so that traditional file management becomes unnecessary?
- Organising folders is old fashioned – haven't you heard of *System X*?

We discuss each of these potential future directions in turn.

That search, tagging, or any other feature will replace or preclude the need for folders and organising one's files is an alluring but likely specious hope for the majority of users. Consider the conclusion of the discussion about search and navigation, above: though desktop search is undoubtedly useful when navigation fails, folders and navigation aid recognition and reminding more than searching, which lets memories become foggy and thus difficult to recall later. Ill-defined information needs are better supported by navigation (or browsing) than by methods requiring the user to explicate that need (Julien et al., 2013) or remember an attribute of an item (e.g., its location or name), and sensemaking is supported by a division of the collection, achieved by folders and reinforced by navigation (Jones et al., 2005). Desktop search is powerful, especially when equipped with full-text indexing, but it lacks the dataset that makes Web searching so powerful (e.g., billions of pages and past queries).

Recent work provides further evidence towards this, finding that over a two week period of attempting to perform FM tasks without navigating their folders, some participants were unable to abstain from using folders, later claiming a dependency and

implicating folders as essential in PIM task execution and the high-level conceptualisation of their collections (Whitham & Cruickshank, 2017). The previously discussed work by Benn et al. (2015) provides clues about why folder arrangement may become so ingrained in user behaviour: the human brain has better built-in support for spatial cognition and recognition than for linguistic processing and recall. The likeliness that searching will replace navigating folders is therefore nicely summarised in the paper title *The perfect search is not enough* (Teevan et al., 2004).

In addition to search, other changes in specific computing contexts to how file management is done may lead some to think FM will soon be obsolete, and possibly therefore something that only power users do, like using the command line. This may be motivated by, for example, software that hides file management from the user in favour of managing at the level of a collection or format, like iTunes, as discussed in the introduction. The thought may go as such: why not forego file management entirely, and instead interact with files only when viewed as digital objects of a certain type, in the applications relevant to each type? This is the paradigm, for example, in Apple's mobile operating system, iOS: applications are *sandboxed*, or restricted to only seeing files they are responsible for.

It is telling that Apple has not implemented this approach in their desktop OS and have added a file manager to the mobile OS. On the desktop, beyond iTunes, the Photos application, and a few other programs files are still interacted with as files, and the Finder file manager application is regularly updated. Arguably, a strict sandbox only moves the general problem of item organisation from the file manager to the format-specific application (e.g., iTunes): items of some format must still be stored, named, organised, assigned metadata, and so on, and once a sufficient number of such items has been stored, it becomes necessary to organise the items with various divisions (i.e., folders or something like them) to facilitate accessing individual items and understanding the whole collection. Interacting with such items without viewing them as files may also be inconvenient; for example, when being sent an email attachment to edit and return, iOS users must download it and hope it appears within the application they intend to use to edit it, and must then push it back to their email from that application. This

may be why Google's Chrome OS, despite encouraging the user to do everything in Web-based applications within the Web browser, has a dedicated, if minimal, file manager application.

The guise of avoiding file management is thus lost once interactions beyond basic access to items are desired: when a user wants to send specific songs or photos to another person or device, they may use Dropbox or a USB connection, and will be sending the items as files. But for users of iTunes, this is not a trivial task: the functionality for synchronisation provided by iTunes is to sync an entire collection, songs' file paths are not created by or familiar to the user, and flexible groups of file paths that would have been created by folders (such as an *Artist X* folder) are not readily available. Interestingly, the previously mentioned study by Whitham and Cruickshank (2017), in which participants failed a focused attempt to stop using their folders, took place exclusively in Mac OS. Sandboxing also entails design choices about file type associations, and these are typically motivated by political and commercial desires rather than usability concerns; for example, Apple's music application on iOS does not, without extensive modification, play FLAC format files, and so users must use another application or convert their FLAC files to Apple's proprietary lossless format.

Sandboxing may also make anything beyond lightweight, casual media consumption challenging. For example, in POSIX systems (Mac OS X, GNU/Linux, Unix, BSD; i.e., everything but Windows), everything is regarded as a file – even drives that read removable media. And so, at least for developers, there are too many digital items to not have some abstraction for interacting with, sorting, and accessing them.

Thus, the file and folder paradigm is not easily replaced: there is a need for a common method for interacting with digital items and for organising those items. Sandboxing seems to avoid some of the entailed difficulties of FM, and does so cleverly by drawing on rich file metadata (the sandbox approach works much better for music in standardised formats than for documents), but creates problems of its own for both user interaction (e.g., pushing content) and PIM (e.g., greater fragmentation of a project's files because of their differing formats). The file is a fundamental *cohering* concept between engineers

and users that provides a common method for interacting with digital content, and thus "remains central to systems architecture and to the concerns of users" (Harper et al., 2013, p. 1125). Improving upon it therefore likely requires incremental change rather than abandonment: "new abstractions are needed, ones which reflect what users seek to do with their digital data" (Harper et al., 2013, p. 1125).

Finally, in light of the problems identified above in using FMs, it is reasonable to think that a revolutionary idea may be desirable for changing how we interact with digital content. As early as the 1960's, this was the mission of the controversial Project Xanadu (Nelson, 1965), which aimed to avoid the paper metaphor in representing digital content, and incidentally was also the first hypertext system, pre-dating the Web (Nelson, 1965). The original aim of Project Xanadu was to "make a file for writers and scientists, much like the personal side of Bush's Memex, that would do the things such people need with the richness they would want... [via] a simple and generalised building-block structure, user-oriented and wholly general-purpose" (Nelson, 1965, p. 84). Guided by 17 rules, documents in the Xanadu model contain any kind of digital content (precluding the need for *files* as such), are linked to other documents based on similar content, and are intended to be edited while being compared with such items; this is meant to utilise the digital nature of the documents to support non-sequential authoring and minimise writing efforts being doubled across documents. Project Xanadu has proven to be as complex as it is promising, and is still in development. It therefore remains unclear how the average user, struggling to meet the challenges of classical FM, would feel about using a *Xanalogical* (Nelson, 2000) interface.

In summary, several incremental and revolutionary prospects promise to change the nature of file management, but given many digital items, some functionality for understanding, interacting with, and organising them is needed, and files and folders fulfil this need.

1.5 Conclusion

File management is a ubiquitous and challenging activity. In this chapter we have synthesised disparate works examining this activity, and have identified that such work typically aims to understand users' FM behaviour, the factors determining it, and how these results can be used to improve the relevant systems and services. These studies have been performed by researchers working in information science, personal information management, human-computer interaction, computer science, and so on, and have drawn upon various methods from these fields; the study of FM is thus interdisciplinary and potentially highly impactful for these fields and those with overlapping interests, such as psychology and information visualisation, retrieval, and organisation. This is perhaps unsurprising, given how the apparent fundamental nature of the file and folder context, where users manage items in bespoke information structures.

What the study of FM faces in the future is a daunting, shifting landscape where user behaviour is difficult to study, analyse, and support, because it is nuanced, private, personal, and changing along with its technological context. The implications of increases in use of the cloud, available storage space, fragmentation of information across devices, and complex social information management on FM are unclear. Robust data, data collection tools, models, and theory will be needed to understand and support user behaviour, alleviate common challenges, develop useful software and services, and to make the fascinating behavioural, psychological, and technological findings of FM studies useful to research into other information behaviour and structures.

Transition 1

In the previous chapter I reviewed over 200 works and demarcated a single topic of study to which they belong, *file management research*, and examined their common motivations, methods, limitations, knowledge, and knowledge gaps. In particular, I identified the need for a broad and confident quantitative description of typical file management (FM) behaviour to enable more advanced study of FM, and found that incommensurabilities in past studies have so far prevented such a description from emerging. I discussed that such a description can not currently be derived as the existing quantitative data collection tools do not collect the necessary data and entail difficulties in administration and recruitment, particularly by requiring a great deal of researchers' time and presence to oversee and by being unappealing to potential participants. In the next chapter, I describe developing, testing, revising, and sharing data collection software to treat these issues, and discuss how it may be used in additional FM research.

Chapter 2

Cardinal: novel software for studying file management behaviour

Abstract

In this chapter we describe the design and trial use of Cardinal, novel software that overcomes the limitations of existing data collection tools used in personal information management (PIM) studies focusing on quantitative descriptions of file management (FM) behaviour. Cardinal facilitates large-scale collection of FM behaviour data along an extensive list of file system properties and additional relevant dimensions (e.g., demographic, software and hardware, etc.). It enables anonymous, remote, and asynchronous participation across the 3 major operating systems, uses a simple interface, and provides value to participants by presenting a summary of their file and folder collections. In a 15-day trial implementation, Cardinal examined over 2.3 million files across 46 unsupervised participants. To test its adaptability we extended it to also collect psychological questionnaire responses and technological data from each participant. Participation sessions took an average of just over 10 minutes to complete, and participants reported positive impressions of their interactions. Following the pilot, we revised Cardinal to further decrease participation time and improve the user interface. Our tests suggest that Cardinal is a viable tool for FM research, and so we have made its source freely available to the PIM community.

2.1 Introduction

Every day, computer users interact with files and folders, including creating, downloading, naming, moving, saving, copying, reviewing, navigating, searching, and deleting them. This is a deeply personal and psychological activity (Lansdale, 1988) that can be supported by systems and services, but such support requires understanding the behaviour that users exhibit and the factors that influence them. Despite many studies of Personal Information Management (PIM) reporting on how people perform file management (FM), a confident quantitative characterisation of FM behaviour has not emerged across existing studies. Such a description would be useful for enabling additional advanced research methods in FM, like principal component analysis, user modelling, and the generation of PIM theory, but cannot currently be collected due to limitations in the available data collection methods. Here we introduce Cardinal, software that addresses these limitations by automating the mass collection of quantitative data about FM behaviour, thus facilitating a detailed quantitative description of users' FM behaviour to complement existing qualitative descriptions. In what follows we describe the existing FM data collection methods and need for Cardinal, detail its design, report on a trial implementation, and conclude by noting the remaining improvements that may benefit FM research.

2.2 Problem area

Broadly, PIM is an area of study concerned with how and why individuals manage information items, and how the results of these investigations might be used to improve services and systems designed to support such management. Understanding FM behaviour and its factors aids the design of PIM systems and services, for example by revealing user preference and behaviour in certain contexts. In time, such improvements are implemented in widely used software and improve the FM experience; desktop search and file tagging are examples of this process, having been developed and tested in academic and industrial research before being implemented into major operating systems (Kljun et al., 2015b). An extensive review of FM literature is provided in Chapter 1.

Many PIM studies have examined aspects of FM behaviour, for example how people name (Carroll, 1982) and organise (Hardof-Jaffe et al., 2009b) files, the challenges of information fragmentation across multiple devices (Capra, 2009), and the various challenges to sharing and retrieving files (Bergman, Whittaker, & Falk, 2014). Together, such studies have provided some characterisation of users' FM behaviour, extending characterisations of user's paper-based organisation strategies into the digital domain and advancing the characterisations from neat or messy and using files or piles (Malone, 1983) to include mixed approaches (Trullemans & Signer, 2014a) and strategies such as filing the majority of files on creation, filing somewhat extensively but leaving many items unfiled, or filing occasionally but leaving most files unfiled (Boardman & Sasse, 2004). At least three methods are used to collect data about FM behaviour:

1. ask participants about their FM behaviour, for example in a questionnaire or interview

2. observe the behaviour directly, for example in a *guided tour* of the desktop or during an experimental task, and

3. infer the behaviour from properties of the file system, for example by running software on participants' computers.

Each method entails benefits and limitations, and the three may complement each other when used in conjunction to answer particular research questions. Though this chapter is concerned primarily with the third (i.e., inferring user behaviour by examining the file system), brief summaries of the first two are provided before the third is discussed in depth.

The first approach, asking participants, utilises an established and rich tradition of ethnographic enquiry, and has the benefits of being relatively simple and direct, as data about PIM-relevant perceptions and behaviour can be reported by participants, for example when elicited in an interview (Xie et al., 2015). This works well for identifying broad PIM practices and challenges that users remember, like transferring files between

computers (Capra, 2009). It is limited, however, as it cannot capture data about activities or aspects of behaviour of which users may not be cognisant, like the number of empty folders they keep, and participants' perceptions of their own PIM behaviour can be inaccurate (Bergman, Gradovitch, et al., 2013b).

The second approach, observing participant behaviour, entails recording participant behaviour, for example using video to capture the behaviour exhibited during typical work tasks (Bruce, Jones, & Dumais, 2004), guided tours of the participants' desktops (Barreau, 1995), or structured experiment tasks (Bergman et al., 2012; Benn et al., 2015). This allows for exploring particular aspects of user behaviour in depth, like organising downloaded files (Jones et al., 2001) and retrieving shared files (Bergman, Whittaker, & Falk, 2014). The limitations of this approach are its temporality and impracticality: as the behaviour is always observed during some particular time, the researchers necessarily do not see what participants are doing when not observed and meeting with participants for guided tours or reviewing recordings of experiments are both very time and labour intensive.

The third approach, heavily utilised in much of the FM research literature (as discussed and summarised in Chapter 1), is to infer and understand users' behaviour by examining the file system, its contents, and its properties, typically by running custom-made software on participants' computers. User behaviour determines properties of the file system (e.g., the shape of the folder tree structure, the particular file system contents, the size of the collection), and the file system therefore serves as a record of such behaviour. For example, recording folder names allows for discerning if particular conventions are used when a user names folders. Properties of particular files and folders can also be analysed together to ascertain subtler facts about user behaviour, such as the average depth at which a user stores document files or the number of files stored in folders that contain no sub-folders. Studies using this approach have, for example, examined the number and kinds of files people store (Gonçalves & Jorge, 2003b), explored how files are organised across folders (Khoo et al., 2007), sought to determine the effect of personality style on desktop tidiness (Massey et al., 2014), observed files' sizes (A. J. Smith, 1981)

and names (Carroll, 1982), explored the role of provenance in file retrieval (Jensen et al., 2010), studied the document management behaviour of students (Henderson & Srinivasan, 2011), and examined the effect of folder depth on file retrieval (Bergman et al., 2012).

Examining the file system has clear interest in PIM research, and the many studies utilising the approach attest to the value of quantitative descriptions of users' behaviour for forming a complete understanding of relevant phenomena (i.e., complementing qualitative descriptions). But the many quantitative descriptions of FM have emerged from studies like those described above, with differing goals and research questions. They therefore naturally feature differing populations (e.g., academics or engineers), differing contexts (e.g., work files or personal files, Windows or Mac OS), and differing measures of the file system (e.g., average file depth or length of file names).

Thus, across the existing studies, no singular broad and extensive quantitative description of FM behaviour emerges or is deducible from analysing the findings of such studies together (e.g., meta-analysis): despite there being at least 40 file system properties available (discussed below), we find that studies often report different measures – e.g., the number of files left in root folders (Henderson & Srinivasan, 2011) or the depths of leaf folders (Zhang & Hu, 2014) – while some measures are never examined, and among 37 previous studies it is common to feature fewer than 5 properties (e.g., Boardman & Sasse, 2004; Henderson, 2005), with the mean being 4.4 and the maximum being 13 (e.g., Gonçalves & Jorge, 2003b). This patchwork of quantitative descriptions is due to differing research goals, rather than researcher oversight, but nevertheless results in a limited quantitative description of general FM behaviour that is insufficient for supporting advanced research methods like principal component analysis and detailed user modelling.

It is currently impossible to collect the data to provide a quantitative description like the one described above. This is not only because existing quantitative data collection tools (i.e., those used in previous studies) collect data narrowly, as evidenced by the existing results and discussed above, but because they impose difficulty in administration and

recruitment. This difficulty entails that such tools currently allow collecting data only from limited population samples; for example, the software may require researcher supervision to use, entailing a small sample, or may be an obstacle to recruitment because it is difficult to find users willing to expose and share their digital possessions and desktops, entailing that participants are from an available but niche group like trusting colleagues. When large sample sizes have been achieved in previous studies, they have been from niche contexts that may not be representative of typical FM, such as students using a proprietary, online environment during a class assignment (Hardof-Jaffe et al., 2009a) and employees at a single software corporation (Douceur & Bolosky, 1999; Agrawal et al., 2007).

Existing software has also rarely supported multiple operating systems, and to our knowledge no existing tool supports Windows, Mac, and Linux. Perhaps as a result researchers have instead relied on tools packaged with the OS (as in, for example, Evans & Kuenning, 2002) that provide minimal functionality, and typically focused on a single OS (as in, for example, Khoo et al., 2007). Consequently, suggestions that software factors such as the OS and file manager used have an effect on FM behaviour (e.g., Barreau, 1995; Massey et al., 2014) have gone virtually unexplored.

A broad quantitative description of general FM behaviour is therefore desirable, but cannot currently be derived from existing results, and although such a description can be derived by collecting data about people's file systems, no tool exists for doing so. This must be addressed to produce a more complete picture of how people manage files, necessary to advance PIM research, for example by producing models of user behaviour, theories of PIM, and creating accurate datasets to use when evaluating PIM tools (Chernov et al., 2008). What is needed, then, is software that collects data about many file system properties, including those used in previous studies, to provide a broad description of behaviour, and that software must facilitate rapid and relatively easy collection across a large, heterogeneous population sample.

2.3 Cardinal - design and use

We created software, called Cardinal, to overcome the limitations of the existing quantitative data collection tools used to study FM behaviour. Cardinal is cross-platform (e.g., runs in Windows, Mac OS X, and GNU/Linux), and will run in multiple versions of each OS on computers with both 32- and 64-bit processors. It does not require that users install it, but rather that they download a single small (<30MB) file, for example from a research project's Web site, which can then be run remotely, without researcher supervision, or with supervision, for example in lab settings. Both manually retrieving the data from participants and asking participants to manually send their data are avoided: upon the user's request the resulting data is encrypted, compressed, and sent to a predetermined destination. Cardinal supports sending data to the researchers' computer via secure file transfer protocol (FTP) and to Dropbox via the provided API. Data is stored in the common JSON format so that it can be imported in bulk into statistical software for analysis; an example of the raw data, with added annotations, is presented in Appendix A.

To overcome the limitation of narrow data collection, we programmed Cardinal to collect 27 of the 28 file system properties collected by previous studies, and 11 of 12 additional properties, totalling 38 of 40 possible properties – 25 more than reported in the previous broadest study of FM behaviour. The two excluded properties are discussed in this section, and a summary of all mentioned properties is presented in Table 2.1. Cardinal also collects properties about the technological factors discussed above (e.g., OS and FM software used), and further data may be collected by including additional fields or questionnaires.

Cardinal functions by iterating through the folder tree from user-specified starting points using Python's built-in *os.walk* function. To ensure that no sensitive or identifying data is collected, a list of folders that the participant wants excluded from data collection is consulted at each step and specified folders are noted but ignored instead of examined. Locations outside of those specified (i.e., system folders) are not examined during data collection, nor are hidden folders (e.g., `/home/jesse/.cache/`) or any folders that the

Property category (previous + new)	Previously examined properties (28)	New properties (12)
Storage (11 + 6)	Hard drive capacity, use, and free space; total files, total folders; collection size (in bytes), collection size (files + folders); file extensions/types; file sizes; file age, shortcuts/symlinks	Available drives; folder age; hidden files, hidden folders; duplicate files (by hard link), duplicate folders (by hard link)
Organisation (10 + 2)	Root folders; tree breadth, tree depth; folders in each folder (branching factor), files in each folder; file depths, folder depths; branch consistency or skewness; use of desktop for storage, use of default folders	Inaccessible folders; presence of user-excluded folders
Naming (5 + 2)	File or folder name*, length of name, numbers in names, punctuation or special characters in names, duplication of names	Letters in names, whitespace in names
Retrieval (2 + 2)	File access times, file modify times	Folder access times*, folder modify times

Table 2.1 – 28 file system properties previously collected in FM research and 12 new properties, categorised by relevant phenomena; Cardinal collects 38 of these 40. *By default names are not collected, and the collection of folder access times are not currently collected.

user has does not have unprivileged access to (e.g., `C:\Windows\system32` or `/bin`). For each location visited, Cardinal records the file and folder properties listed in Table 2.1 using the built-in *os.stat* function and other custom functions. For example, *os.stat* returns the size of files and the last time a file or folder was accessed or modified. Folder modify time is a previously unused property that is updated by the OS when the user adds or removes a file or subfolder, or renames the folder; this may be used to better understand how users perform organisational or meta-level PIM activities (Jones, 2007a). Folder access times are not currently recorded because this property is set to the current time by the OS at the moment Cardinal reads the contents of the folder. We plan to address this in a future update to Cardinal by reading the property before accessing the folder.

Since we designed Cardinal to not store file and folder names, semantic measures are calculated and stored as each file and folder is examined, including the previously used properties of name length, use of numbers, use of special characters, and detection of duplication of names, but also records the use of letters and whitespace.

Other new properties collected include identifying files and folders that are hidden or duplicated across multiple hard links. Previous studies have examined how users manage duplicate files and folders as identified by duplicate names (e.g., Hicks et al., 2008; Henderson & Srinivasan, 2009). Files and folders can be duplicated in a number of ways; for example, by making a copy, maintaining two files with the same content and name, or by creating a hard link. Files are themselves hard links to some data on a disk, though additional hard links to that data can be made such that two files really provide access to the same content, or in other words, these two files really are the same file but the use manages its existence across multiple locations. Cardinal identifies when files and folders have been duplicated in this way by checking the *nlink* property returned by *os.stat*; a value greater than 1 entails duplication via multiple hard links to a file.

Hidden items have not been examined in prior PIM research, but may exist in the user's collection as a result of the user unintentionally downloading or explicitly hiding them, and require special attention to manage since FM display settings must be toggled

to view them. To protect user privacy, Cardinal does not record properties about hidden files, nor does it enter hidden folders while collecting data, but it does note the existence and locations of such files and folders.

Files and sub-folders are assigned to folders by ID so that further properties can be derived later, like tree topology (e.g., depth and breadth); in essence, a mirror of the hierarchical arrangement of files and folders is made. This means researchers can later make *post hoc* measures of the mirror that would be impossible to derive from a flat list of files and folders. For example, rather than being limited to mean file size, derived from a list of file sizes, the distribution of file sizes or types across folder depths can be derived by examining the files where they are located across the folder tree.

We validated the accuracy of the data collected by Cardinal in purpose-made testing environments in each compatible version of Windows, Mac OS, and Linux. These environments were folder structures populated with files of varying properties (e.g., age, name length, location), hidden files, symlinks, etc. such that each of the file system properties measured by Cardinal could be manually measured and verified against Cardinal's output.

Once the provided executable is downloaded and run, a simple interface (seen in Figures 2.1-2.4) walks the user through the following steps:

1. Greets the participant, outlines the process, and presents a consent form (Figure 2.1).

2. Asks for demographic information (age, occupation, education, gender) and the form (laptop, desktop, tablet, other) and use (work/school, personal, both) of the computer (Figure 2.2). Responding to these questions is optional, though it can be required in the interface.

3. Asks for the names of installed software relevant to file management and suggests any likely values based on the OS detected (e.g., Finder for Mac, File Explorer for Windows).

4. Asks the participant to select folders that they personally manage, suggesting the

user's home folder as one location.

5. Allows the participant to select folders that they wish to have excluded from data collection.

6. Allows the participant to initiate the examination of the selected folders, collecting file system data while ignoring file contents and file and folder names.

7. Presents any included questionnaires to the participant (Figure 2.3).

8. Presents a summary of their collection and results of any questionnaires, and asks the participant to initiate submitting the collected data to the researchers (Figure 2.4).

9. Thanks the participant and exits the application.

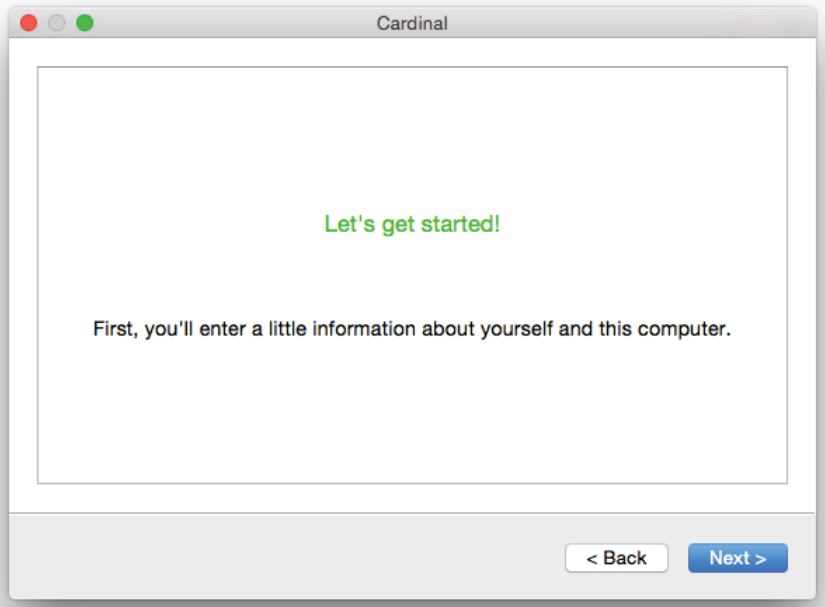


Figure 2.1 – Cardinal's user interface: a sign-post page greeting the user.

To encourage participation, we aimed to make Cardinal simple and easy to use. For example, it appears native on each OS to reduce unfamiliarity, requires little time of participants (specific measurements are presented in the next section), and is laid out

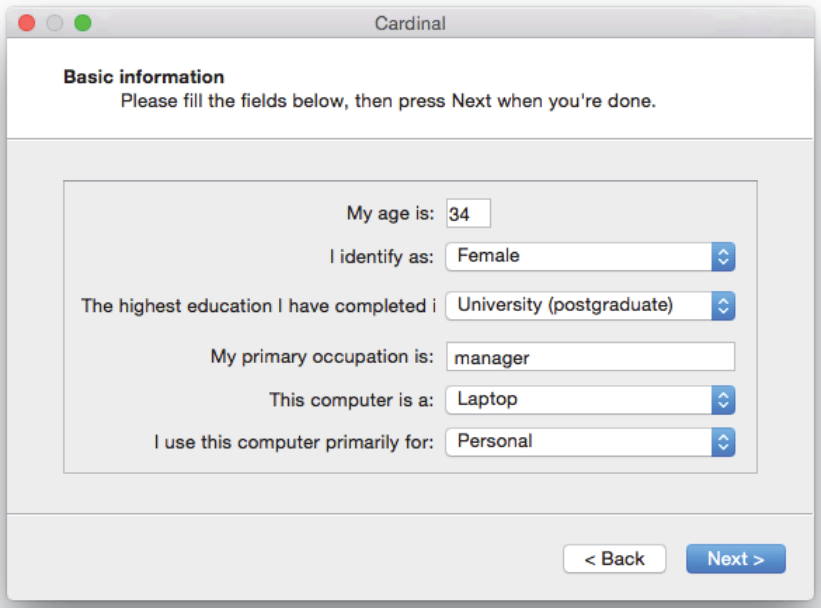


Figure 2.2 – Cardinal's user interface: a page for the user to enter demographic data

sequentially, with back and next buttons and instructions on each panel. During development we employed a simple iterative design process by soliciting free-form feedback from five colleagues through email. Though all five users were able to make basic use of the software, two rushed through the pages without reading the instructions and then expressed feeling confused about what they were meant to do, so we inserted *sign-post* pages containing summaries of what general task comes next.

All five users expressed concerns about privacy that arise from exposing their file collections, so we configured Cardinal to respect participant privacy: participation is anonymous as the identities of the participants are never known to the researchers, sensitive folders can be excluded from data collection, and identifying file and folder properties are respected as described above. Two users still noted feeling unsure about what the software had seen, so we added an instant summary of the results of their participation (e.g., their most common file type, the length of their longest folder name, and the number of empty folders), which they said alleviated their concerns enough to use the software. We also added a link to a Web page displaying averages of the FM data collected thus far so that they could compare their own results. Though providing such a page remains

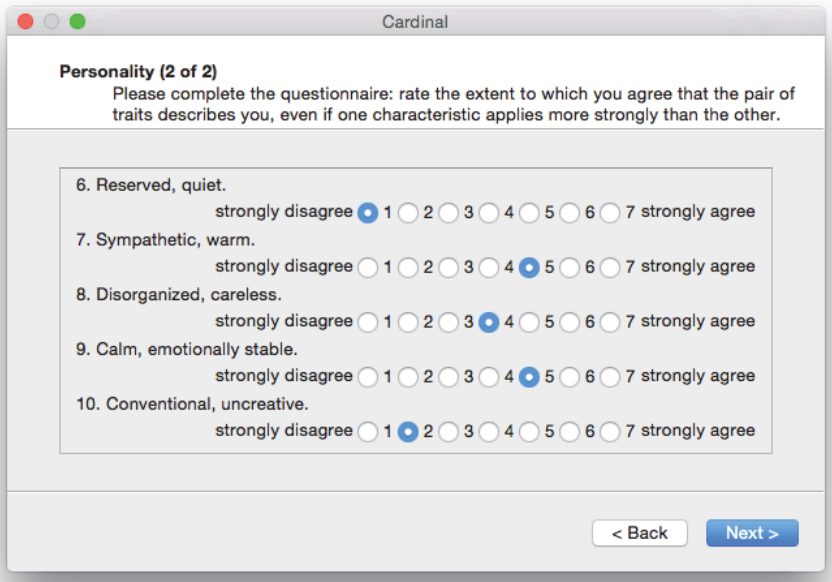


Figure 2.3 – Cardinal's user interface: a page presenting an included questionnaire (example items from Gosling et al. (2003)).

an optional aspect of using Cardinal in future studies, it may encourage participation by making the data collection more transparent and meaningful to potential participants. Cardinal is also open-source software, thus some degree of trustworthiness is implied by the code being visible to a community of developers and open for interested participants to review for themselves.

As the software facilitates rapid distribution, recruitment can be tailored to reach the intended population, and any participants not meeting demographic criteria can be filtered out afterwards. For example, with the software hosted on a Website, traditional recruitment methods (e.g., fliers, emails, social media) may point to the page and participants can help themselves to the software. Direct compensation is made impossible with anonymous participation, but participant identification could be added by including a free text field (e.g. for inputting email addresses), and internal motivation may come from the participants' desire to learn about their own FM behaviour, which is summarised and reported to them at the time of data collection. Improved distribution and recruitment may make the software attractive also to computer science researchers, like those working on file system design and file-size distribution (e.g., Tanenbaum et al., 2006),

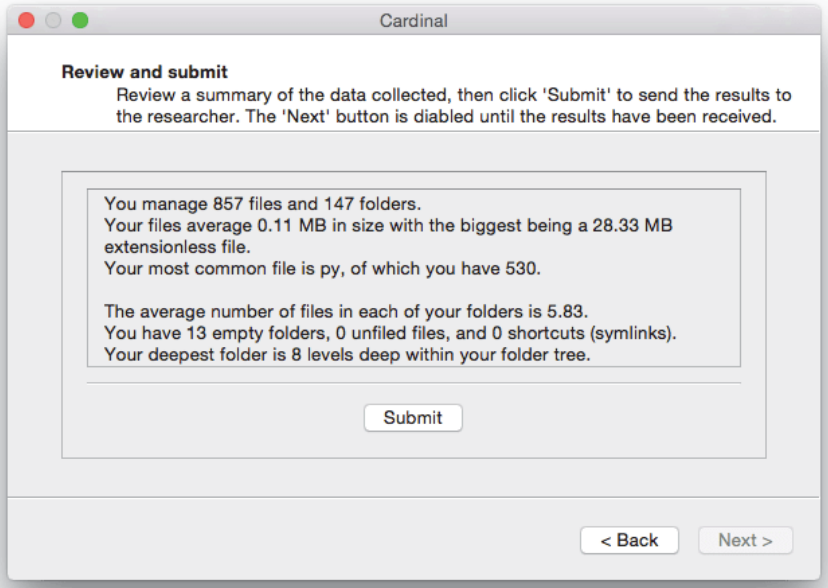


Figure 2.4 – Cardinal's user interface: the results summary page. Further results are viewable by scrolling or enlarging the window.

where small and niche population samples have been a research limitation as much as in PIM research (Douceur & Bolosky, 1999).

To aid reuse in further studies, Cardinal was made using open-source tools[1], and we have shared its source[2] under a liberal license (GPL 3). Next we describe its use in a trial implementation.

2.4 Trial implementation and subsequent improvement

We implemented a pilot study[3] to demonstrate a use case for Cardinal and test its efficacy as a data collection tool. We emailed 48 people (12 faculty and 36 PhD students) in our department and provided a link to a Web page that explained what participation entailed and contained links to download the software. Within 13 days, we received 21 responses

[1]Python 3, the Qt graphics framework, and PyQt bindings
[2]`https://github.com/jddinneen/cardinal`
[3]Our pilot study was approved by the McGill University Research Ethics Board (#75-0715).

(44%). In two following days we invited 82 master's students to participate, and received 25 responses (30%), resulting in 46 of 130 possible participants (35%). Collection was successful on both laptops and desktops running the 3 supported OSes (26 Windows, 19 Mac, and 1 Linux). In total, Cardinal collected data about 2.3 million files and 290 thousand folders, and recorded questionnaire responses and technological data (OS and FM software used) for each participant.

Time stamps were recorded each time a new page of the interface was accessed. Excluding two outliers discussed below, the mean time taken to complete a session was 10.6 minutes (SD = 7; min. 2.5; max. 33.4), of which an average of 7 minutes (66%) were spent reading the consent form, entering demographic data, and answering two questionnaires bundled within the software. The remaining time was passed collecting data about the file and folder collection and preparing a summary of the data. The former took an average of 1.86 (SD= 2.7) minutes, accounting for 17.5% of the time to complete a session, while the latter took an average of 1.69 minutes (15.9% of the completion time).

Participants' use of Cardinal was largely unproblematic: responding to the invitation email, two participants reported that using Cardinal was 'a breeze' and 'painless', and six participants reported finding their summarised results to be of interest, noting for example that they did not know they had so many empty folders or large files. Two issues in using Cardinal were identified during the trial. First, one participant was unsure if they should plug in external hard drives to be examined. This should therefore be clarified in the participation instructions of each study implementing Cardinal. Second, four participants stated that the software appeared unresponsive while collecting and summarising data about large numbers of files. This was solved by putting the relevant processes on a separate processor thread so that the interface stays responsive while they are running.

Given that participation was done remotely and in a potentially wide array of software environments we expected that Cardinal may encounter some errors or fail to run in at least a few cases. Indeed, three participants had Mac OS versions that were too old to run the software at the time of the pilot. To remedy this, we compiled Cardinal in an older Mac OS X version, and it now runs on versions 10.8 and above, supporting 90% of

the Mac OS X market.[4]

The outlying participation times for Cardinal were 1.25 and 12.75 hours. The participant with the longer time emailed us to explain that they left Cardinal running overnight to complete the results summarisation, and analysis of the time stamps revealed that this took nearly all 12 hours of the completion time. This was the longest summary time by approximately 11.5 hours. The lesser outlying completion time was primarily due to 33.9 minutes of file system data collection. This was nearly twice as long as the second longest collection time.

These outliers are extreme and surprising given that neither collection was the largest one seen in our pilot study. Similar cases may arise in future data collection, so we attempted to decrease the time required to perform both the data collection and results summary phases. To speed up the data collection, *os.walk* was augmented with a function called *os.scandir*, which iterates through directories faster. We also revised our approach to generating a summary of the participant's results by deriving several measurements more efficiently.

To understand the impact of these changes, we analysed a test collection consisting of 222,321 files and folders (5% larger than the largest participant collection) using both approaches. Where the old approach, using *os.walk*, took 11.5 minutes to collect data about the test collection and 56.3 minutes to summarise the results, the new approach, using *os.scandir*, took only 1.45 minutes to collect the same data (an 87.4% decrease in time) and the new summarisation approach took just 1 second (less than 0.03% of the original time). This implies an improved data collection time of 4.3 minutes (down from 33.9) for the most outlying collection time, and an improved summary time of 12.9 seconds (down from 12 hours) for the most outlying summary time. Considering these improvements together, we can expect the mean participation time for future participants to be approximately 7 minutes, rather than the 10.6 average seen in the pilot study, so long as the number and contents of questionnaires implemented remain comparable.

[4] `https://www.netmarketshare.com/operating-system-market-share.aspx`

2.5 Limitations

A file management-based approach is only one among several for understanding personal information management. Others may examine physical representations of information, or examine digital organisation but focus closely on cognitive- and context-related aspects. Nonetheless, the approach outlined here of inferring user behaviour by examining quantitative measures of the file system complements these, and has been used in many studies – at least thirty, as discussed in Chapter 1 – to understand users' regular experience of managing, sorting, and navigating their personal information stored in files. Further, file system property data can be used together with other approaches, and to facilitate this we have included provisions in Cardinal for integrating standard cognitive- and context-related instruments. For example, our pilot study included questionnaires related to personality style and spatial cognition, and it would be simple to include other instruments, questionnaires, or free text fields for user-reported data.

Another concern is that inferring FM behaviour from the file system, rather than observing actions as they happen, may capture data about a limited selection of a user's FM behaviour. For example, Cardinal can count the number of folders at the time of scan, but does not indicate if a user created and deleted folders beforehand, nor does it inform us about actions like renaming, moving, or sharing files. In other words, the data produced by Cardinal is a snapshot of a user's file system as it has been produced by their behaviour leading up to any singular point in time. It is desirable to improve upon this limitation, as the importance of longitudinal data will grow as the prevalence of long-term personal information management increases (Dinneen, Odoni, & Julien, 2016; Jones, Bellotti, et al., 2016). This may be partially overcome, however, by repeated executions of the software by the same participant; the data would then together be longitudinal and could be analysed as such.

Finally, since the default setting in Cardinal is to respect participant privacy by not recording file and folder names, the semantic analysis that can follow is limited to the specific properties measured during data collection: name length, number of letters, numbers, whitespaces, and special characters, and name duplication. This necessarily

means that it will be difficult or impossible to identify naming conventions or understand the use of a folder based on its name. This is the price of participant privacy; though Cardinal may be modified to overcome this, it will likely make recruitment more difficult.

2.6 Conclusion

We have developed Cardinal to overcome the limitations of the quantitative FM data collection tools used in PIM research, specifically: narrow data collection, impractical administration and recruitment, and supporting few OSes. In a trial implementation of 15 days, Cardinal collected FM behaviour data along 38 file system properties and additional demographic and psychological data from 46 participants and did so remotely, asynchronously, and across three OSes. This indicates it is a viable tool for collecting quantitative data about FM behaviour, and is an improvement over the previous similar data collection tools because it eases administration, collects nearly all kinds of previous file system metadata and new ones, should scale well to facilitate longer collection periods over larger and more heterogeneous samples, and has been shared for reuse.

Cardinal can therefore be instrumental in facilitating an increased understanding of FM behaviour, which has the potential to enable future FM research towards identifying the principal components of FM behaviour, modelling users' behaviour, advancing PIM theory, and aiding the design of future PIM systems and services. Cardinal can also be used in studies of FM behaviour where measures of the file system provide complementary data; for example, file depths and access times collected by Cardinal could be used in analysing the results of an experiment using prompted file retrieval tasks. We are happy to share Cardinal with other researchers and hope it will save time and effort in future PIM studies.

Transition 2

In the previous chapter I described the design, testing, revision, sharing, and potential uses of quantitative data collection software intended to treat the knowledge gaps and methodological limitations identified in the first chapter. In the next chapter I report on a study making use of that software to provide a broad, quantitative description of typical FM behaviour, the need for which is identified and discussed in the first chapter and discussed further in the second chapter.

Chapter 3

Growing collections, stable organisation: an extensive quantitative description of how people manage files and folders

Abstract

File management (FM) is a ubiquitous computing task and common in personal information management (PIM). Designing software and services to support FM requires a detailed understanding of users' relevant behaviour, motivations, and challenges. While numerous studies have focused on the latter two, an extensive quantitative description of typical user behaviour has not been established, but is needed for a more complete and advanced understanding of user behaviour capable of generating models of FM and a theory of PIM. In this chapter we provide such a description by examining 56 measures of the FM behaviour exhibited by 301 participants as evidenced by snapshots of their file systems. Despite omens that files are outdated and will be replaced by desktop search or tagging, we found that users are engaging in *more* file management by keeping considerably larger collections than previously observed and storing these in taller and wider folder trees. Despite the growth, these trees' internal structure and file categorisation has remained stable and hard drives remain half full on average. We also found that data along most measures were log-normally distributed, requiring special analyses to characterise accurately, and indicating that previous studies have underestimated typical values and their ranges. We discuss additional findings and implications, and make suggestions for the next steps towards an advanced understanding of FM behaviour.

3.1 Introduction

File management (FM) is a ubiquitous computing task wherein computer users create, name, rename, download, move, copy, organise, delete, share, navigate to, and search for digital files and folders. Many studies of personal information management (PIM) have advanced an understanding of and ability to support FM by examining users' needs and common decisions and challenges, for example by conducting guided tours of users' desktops, observing users during solicited file retrieval tasks, or designing and testing novel system improvements. To further advance scholarly knowledge about FM and better support users undertaking the activity, it is desirable to be able to identify the principal components of user behaviour, model such behaviour, and determine the relative importance of its various determining factors. Requisite for such work is a detailed description – both qualitative and quantitative– of users' behaviour that is rich in context, scope, and descriptive power, but a broad quantitative description of FM behaviour has so far been absent from FM research.

This chapter reports on a study attempting to derive such a description by examining the FM behaviour of 301 participants using Windows, Mac, and GNU/Linux, as evidenced by 27 of the 28 previously reported properties of the file system and 11 additional properties not examined in previous research. The resulting data are analysed to produce 56 measures of FM behaviour, including their file and folder storage and naming, folder organisation and file categorisation, and file and folder access behaviour. This enables comparisons and synthesis with the findings of previous studies and provides a broad, confident, and current quantitative description of typical FM behaviour. In what follows we briefly review the problem space by summarising the gap in PIM knowledge identified in previous chapters, detail our methodology and analyses, discuss our results and compare them to previous studies, note the limitations to our approach, and draw conclusions from our study.

3.2 Literature review and research questions

In this section we characterise personal information management (PIM) and file management (FM) and outline the problem motivating our study. We define the scope of PIM broadly as any instance of personally managing information, regardless of whether the information is about or is owned by the person managing it. This includes, for example, a knowledge worker managing project files on their company computer, a parent maintaining the family calendar, a researcher updating their references list, a student clearing their email inbox, and someone searching for a memento in their personal archive or collecting information about their family history. PIM is therefore quite broad, but its common activities can and have been categorised into fundamental groups like storing, organising (also called meta-level; e.g., including structuring the folder tree and categorising items within it), and retrieving (Jones, Dinneen, Capra, Pérez-Quiñones, & Diekema, 2017) or alternatively, storing, organising, and exploiting (Whittaker, 2011).

FM is thus a specific kind of PIM that involves digital files and folders, and actions performed towards FM can also be described with the three PIM categories; for example, one stores files on their computer, organises them into folders, and retrieves them later. Figure 3.1 shows examples of two representative views of files as would be seen during file management, providing a visual summary of the file management context; a full review of FM research is provided in Chapter 1, covering the motivations, methods, findings, and remaining knowledge gaps of hundreds of relevant studies. The findings of that review that are relevant to the present study can be summarised as:

- While many qualitative studies of FM have helped to identify some common challenges, motivations, and basic kinds of possible interactions pertinent to FM, an extensive quantified description of users' behaviour has not yet been produced.
- A broad description of FM does not emerge from previous quantitative studies because their results are often incommensurable; due to differing goals and research questions, studies report disparate measures of behaviour, feature small or niche population samples, and have incomparable study contexts (examples in Table 3.1).

- Such a description can inform and complement qualitative descriptions of FM behaviour to form a more complete picture of FM, and is needed to undertake advanced research like identifying principal components of user behaviour, modelling that behaviour, generating a standardised and representative file collection for evaluating PIM software, and investigating the factors that determine users' behaviour.

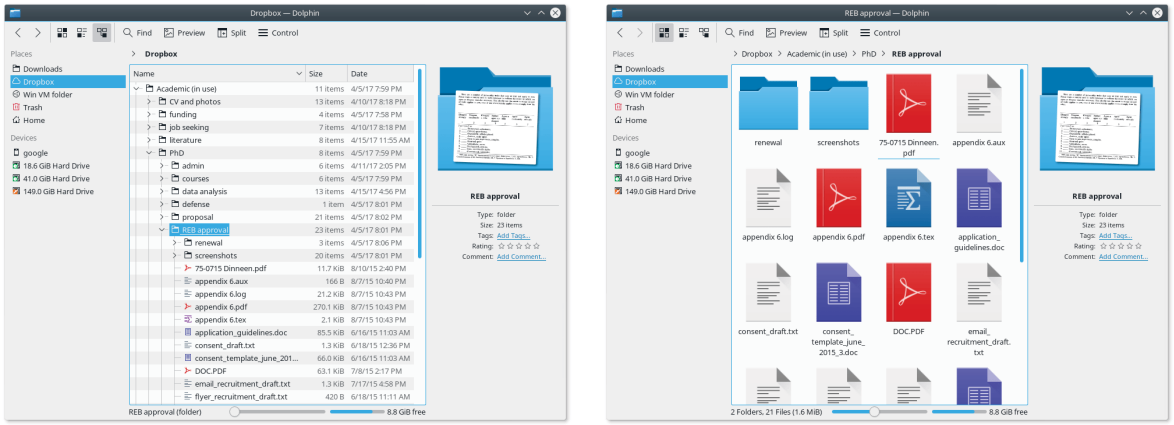


Figure 3.1 – Examples of views onto files and folders, as seen during file management. These may be called *tree* (left) and *icon* (right) views.

study	participants	measures	OS
Gonçalves and Jorge (2003b)	11 academics, professionals	14	Windows, Linux, Solaris
Khoo et al. (2007)	12 professionals	6	Windows
Hicks et al. (2008)	40 engineers	10	Windows
Henderson and Srinivasan (2009)	73 univeristy employees	8	Windows
Whitham and Cruickshank (2017)	12 academics	6	Mac OS

Table 3.1 – Example works inferring FM behaviour from participants' file systems, with participant count and makeup, number of measures made, and operating systems (OS) included. Due to differing study goals and research questions, incommensurability in the quantitative results emerge from niche samples, disparate measures (none are common across all five studies), and narrow contexts.

Put plainly, we can better understand *why* users do what they do and *how* to best support it when we have greater knowledge about *what* they do. This reflects the perspective advanced by Bergman (2013, p. 465): "As PIM research moves from an infant stage of exploratory studies to more rigorous quantitative ones, there is a need to identify and map variables that characterise and account for the variety of PIM behaviour." With

the present study we thus aim to advance the state of PIM research by using novel data collection methods and statistical analyses (described below) that can complement those used previously by providing a quantitative and extensive description of FM behaviour. Therefore the research questions of the present study are:

1. What currently constitutes typical file management (i.e., what are the current values of quantitative measures of FM behaviour)?

2. How do our results differ from those reported along common measures made in previous studies (i.e., when it is possible to see a change in typical FM behaviour over time, has it changed and how?)

Next we outline the methodology used and data analyses performed in our study.

3.3 Methodology

To overcome the current obstacles to a broad quantitative description of FM identified above (e.g., narrow data collection, niche population samples, and overall incommensurability) we aimed in our study to provide a broad, detailed, quantitative description of users' FM behaviour by observing many measures of this from a large and heterogeneous population sample. To produce this description we created and used software, described in Chapter 2, which observes the locations in participants' file-folder hierarchy where they manage files and records thirty-eight file system properties, such as locations and descriptions of each file and folder, and notes relevant information about the present hardware and software. Here we describe our sampling and data collection techniques and the data preparation, classification, and analyses. The study described in this chapter was approved by the McGill University Research Ethics Board (#75-0715).

3.3.1 Population sample and data collection

Participants downloaded the software described in Chapter 2 from our Website and ran it on their personal and work computers. This consisted of answering questionnaires and

specifying through a simplistic graphical interface where they manage files (encouraging the inclusion of both active, working areas and backup locations like external drives), reviewing a summary of the results, and choosing to let the software return the data to the researchers. Participation was therefore remote and anonymous.

Participants' home folders were listed by default as one of several possible locations where they manage files, and participants were encouraged to add any additional such locations. Locations outside of those specified (i.e., system folders) were not examined during data collection, nor were hidden folders (e.g., `/home/jesse/.cache/`) or any folders that the user has does not have unprivileged access to (e.g., `C:\Windows\system32` or `/bin`). Any visible and accessible folders and files within such spaces were included in data collection. This approach accords with the definition of *user-managed files* adopted in this thesis: files and folders that the user stores (either explicitly, by downloading, or implicitly, by not deleting), organises (by arranging, leaving arranged, or not arranging), and retrieves (by any method) in the normally-accessible user space. Examples of such files include those on the Desktop, in My Documents, in Downloads, and any other folders added to the user's home folder.

As our study focused on general computer users managing files, our criteria for participation were only that participants have work or personal files that they manage, and have the abilities to read English and download and run the software. This allowed us to recruit broadly, which we did actively from February to August of 2016, and we continued to passively receive submissions until February 2017. We posted calls for participation on the recruitment Website `www.callforparticipants.com`, posted in recreational online communities (e.g., on Reddit in r/samplesize, r/mac, r/linux, and in Facebook groups), sent emails to the mailing lists of several universities, emailed colleagues, friends, and family, promoted our recruitment efforts at conferences, and invited participants whenever our study came up in conversation.

As the population *frame* in question – general computer users – is indeterminable in size and presumably constantly changing, purposive sampling strategies like stratified probability sampling were impractical for this study. However, the recruitment approach

used facilitates a sample size that is relatively large in comparison to previous comparable studies and reduces the likelihood of a homogeneous population sample (e.g., PhD students in a single department).

3.3.2 Data preparation and FM behaviour measures

The collected data are text files containing descriptions, in hierarchical JSON format, of users' hardware and the portions of their file systems they marked as personally managed – described to the participants as 'locations where you manage files' – including any temporary or *working* folders and those in non-working areas like any external drives they may have nominated. We analysed each received data file using custom Python scripts, which produced many measures of each participant's collections across which we then derived fifty-six measures of FM behaviour.

Table 3.2 summarises the measures produced and how they were gleaned from the participants' data, and categorises them into groups aligning roughly with the commonly used categorisations provided by Jones et al. (2017) and Whittaker (2011): storing (including naming), organising (or meta-level), and retrieving (or exploiting). The categories in this table are later used to structure the presentation and discussion of the results, below. These categories also cover each of the external, immediately observable categories of PIM variables identified by Bergman (2013, organisation, structure, retrieval), with differences in the variables examined: we provide many more measures, but our measures of retrieval differ since we did not observe participants directly, and we provide no semantic analysis of users' file and folder naming. Details of how each file system property is recorded can be found in Cardinal's annotated source code.[1]

As described in Chapter 2, our data collection software enabled making observations of participants' files and folders that have not been made previously. How these and other observations have been used to derive the FM measures above is self-explanatory in only some cases (e.g., total number of files), so we clarify here those that are not.

[1] `http://www.github.com/jddinneen/cardinal`

category	measure group	measures
storage	hardware (1-5)	number of installed drives; capacity, space used, and free space (GB); collection size (GB)
	files, folders (6-17)	collection size (all items); total files, total folders; mean file size (in MB); mean file and folder ages (in days, Windows only); number of hidden files and folders; number and percentage of shortcuts; number of hard links
	file and folder naming (18-31)	mean length of file and folder names; mean number of letters, numbers, special characters, and whitespace in files and folders; number and percentage of duplicated names
organisation	structure (32-45)	number of roots; maximum and mean breadth of folder tree; number of folders at root; number of leaf folders; branching factor; number of switch folders; maximum tree mean depth, waist depth; mean depths of all folders, leaf folders, and switch folders; percentage of leaf and switch folders to all folders
	categorisation (46-53)	files at waist; mean files per folder; number of empty folders, percentage of folders empty; mean depth of files; number of unfiled files, root pile rate; mean depth of files, depth of file waist
retrieval	(54-56)	mean time since files have been in accessed (in days), mean time since files and folders have been modified (respectively, in days)

Table 3.2 – 56 measures of FM behaviour: five hardware measures (1-5) and fifty-one measures of the file system (6-56) categorised by relevant PIM behaviour. These are derived from the thirty-eight hardware and file system properties observed by the data collection software.

Storage measurements We recorded file and folder ages by noting the difference between their creation times and the time at scanning, but provide these measurements only for Windows data since files' creation time metadata is used non-uniformly in Mac OS and Linux. Participants' hidden files were counted, but since participants are either actively hiding them from others or are unaware of their existence, they were not examined (e.g., for their file size or name length, per below) nor counted towards the total number of files. Similarly, we counted pointers to files (e.g., shortcuts, aliases, or symlinks), but again did not examine them nor count them as files. We also counted instances of files or folders being placed more than once into the file system tree, called hard links. File and folder names were discarded, but measurements of them and instances of their duplication were recorded by the data collection software.

We judged the age of participants' collections by subtracting the time of participation from each of the creation times of their files and folders. These measurements should be interpreted cautiously, as programs may over-write these metadata without informing the user (Douceur & Bolosky, 1999; Agrawal et al., 2007), and file creation times may reflect the date the file is moved to a new drive. Because creation-time metadata is changed less frequently in Windows than in POSIX systems (e.g., Linux and Mac OS, where it can reflect the last time of file metadata modification), we report only values observed among Windows participants.

Organisation measurements We describe organisation as the structure or layout of the tree's folders, and as the categorisation of files within it. For the former, we defined roots as user-specified locations from which data collection began recursively, so long as locations were not within one another; at least one was required to collect data, and a maximum of four were allowed by our software. Tree topography was produced by noting the maximum depth (or alternatively, height) of each tree as defined by the maximum depth of a folder, the breadth of tree as the number of folders at the most common folder depth, or *waist* (Hicks et al., 2008), and the waist depth as the most common folder depth. We measured mean breadth as the total number of folders divided by the tree depth, or in other words, the average number of folders present at any given depth.

Leaves are folders containing no folders, thus forming the bottom of a tree, which is not typically even (i.e., not found at one single depth).

We characterise the internal structure of trees by their branching factor. Although there are various ways to define this measure, each used in previous studies, because our other measures already indicate the maximum and average breadth of the tree (i.e., its *bushiness*) we follow studies that define it as the mean number of folders per non-leaf, which indicates the average navigation decision seen in any given folder (Bergman et al., 2010; Henderson & Srinivasan, 2009; Zhang & Hu, 2014). Following Akin et al. (1987) we defined folders as *switch folders* if they contain no files and at least one folder, and counted instances of these.

We made several measures of file categorisation. We defined unfiled files as occurrences of files at roots (typically the participant's home folder or drive root) and therefore root *pile rate* as the ratio of unfiled files to total files, as was done by Hardof-Jaffe et al. (2009a). Our data collection software suggested participants' home folders as one of potentially four locations in which they manage files. We recommended the user's home folder since in Mac and Linux it is the folder closest to the system root that unprivileged users have access to, and in Windows it is the highest folder intended for users' storage. Participants could change this suggested location and add additional locations, and if such locations were *within* other added locations they were not treated as a root and their immediate files were *not* treated as unfiled. Therefore, we did not count files on the desktop, in My Documents folder, etc. as unfiled unless the user defined them as roots (i.e., they explicitly removed their home folder from the specified locations list and added such locations separately).

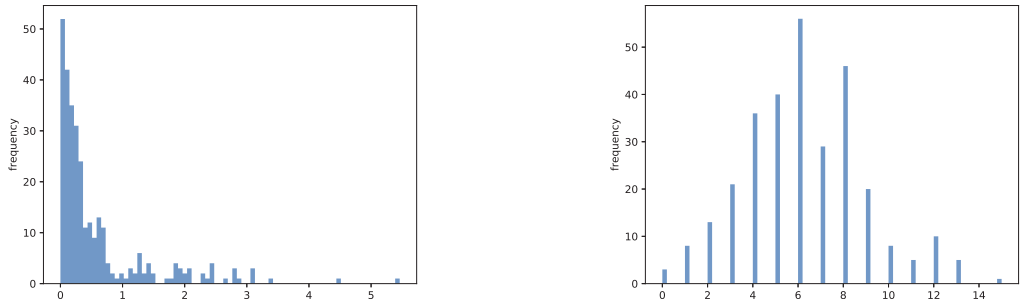
We also recorded the number of files in each folder and files' depths, which allowed us to define a *file waist* as the tree depth with the greatest number of files (i.e., mode of file depth).

Retrieval measurements We defined the time since files and folders were modified, and files accessed, equivalently to how we derive file age, described above. Folder access times were not recorded since they were updated at the moment the data collection software

examined them.

3.3.3 Data classification and analysis

Visual and numeric examination of the data (e.g., with plots and tables, respectively) revealed that the data for most measures were either distributed normally or with long tails (see a comparison of the two distributions in Figure 3.2). Long-tailed data are visually distinguishable from normally-distributed data by, for example, the shape of their histograms, which feature a group of relatively low values and a long *tail* of infrequent but very high values, while normally-distributed data form a relatively symmetrical *bell* shape. This difference is reflected quantitatively in a large skew and measures of dispersion (e.g., standard deviation, SD, and inter-quartile range, IQR) greater than the measures of central tendency (e.g., mean and median); for example, the long-tailed data (a) shown in Figure 3.2 have a mean of 0.57 and SD of 0.79, while the normally-distributed data (b) have a mean of 6.19 and SD of 2.71.



(a) long-tailed, right-skewed distribution **(b)** normal distribution

Figure 3.2 – A comparison of long-tailed (a) and normally-distributed (b) data composed of the same number of data points.

SD values higher than the corresponding means and long-tailed, highly skewed data have been reported in previous studies of FM behaviour (Bergman et al., 2010; Henderson & Srinivasan, 2009; Hicks et al., 2008; Massey et al., 2014), but the data are typically analysed and described as though they were normally distributed. Traditional descriptive statistics derived under the assumption of a normal distribution do not accurately describe skewed, long-tailed data (Limpert, Stahel, & Abbt, 2001), and so we took two different

approaches for describing our normally-distributed and long-tailed data. For normally distributed data, we removed outliers according to the interquartile range method (i.e., removing values that are greater than 1.5 times the third quartile or less than 1.5 times the first quartile)(Wilcox, 2011), and report the arithmetic mean, standard deviation, skew, median, and interquartile range.

To determine which long-tailed distributions our data followed, and thus which descriptive statistics should be reported, we compared possible fits visually and with statistical software.We observed that for every long-tailed measure a *log-normal* distribution was greatly preferable[2] when compared with other long-tailed distributions such as power law and exponential distributions, and was as good or better than the negative binomial distribution.

Log-normal distributions are common in empirical data, and there are standardised measures of their central tendency and spread that are more accurate and informative than those that would result from assuming the data are normally distributed and measuring them in traditional ways (Limpert et al., 2001); see Figure 3.3 for a comparison, where the appropriate measures better reflect the spread and bounds of the range of typical values in a log-normal distribution of data. For our long-tailed data, therefore, we report a median and standard deviation derived by log-transforming, making traditional mean and SD measures, and back-transforming the data (i.e., e^{μ} and e^{σ}, respectively), and we also report a standard *log-normal mean*, defined as $\ln(\mu) = e^{\mu+\sigma^2/2}$ (Limpert et al., 2001; Parkin & Robinson, 1992). Because these are derived differently than normal median, SD, and mean values, we adopt a similar labelling to that used by Limpert et al. (2001) and refer to them as the median*, SD*, and mean*. We also report the number of outliers removed using IQR, as we did for the normally-distributed data, since it does not require symmetrical distribution (Seo, 2006).

Log-normally distributed data are right-skewed by definition (i.e., the majority of the data points fall to the left side of the distribution), and thus have a *range* of typical

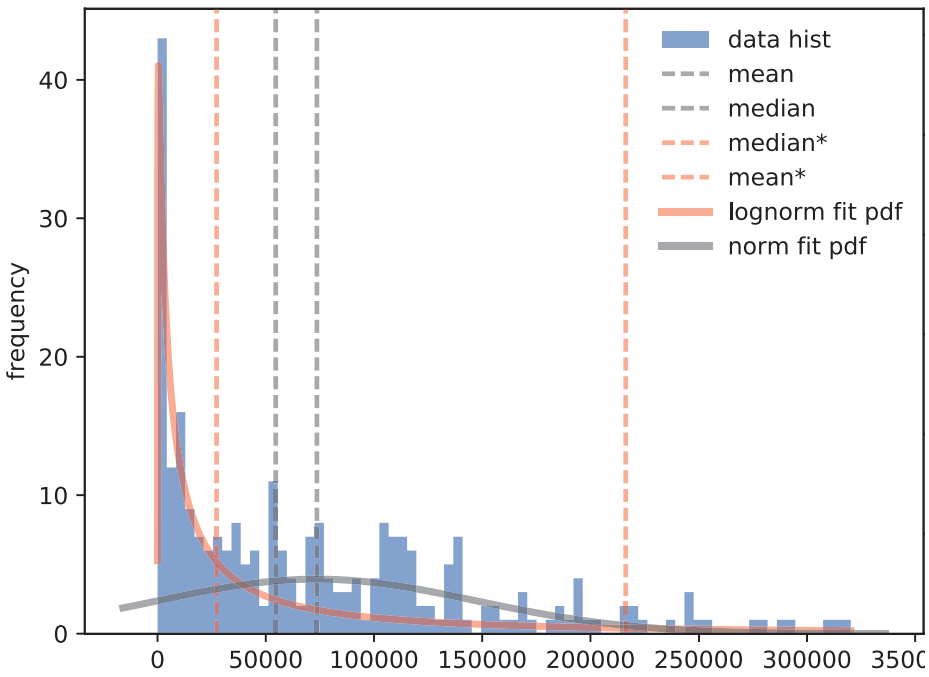
[2]By *greatly preferable* we mean that the fitting package reported a strong preference score for a log-normal fit ($r > 100.0$) when compared to any other fit, and that score was not likely due to chance ($p < 0.05$).

values, the lower and upper bounds of which are indicated by the median* and mean*, respectively. For lightly-skewed log-normal distributions these bounds will be near to each other and will be similar to their traditional counterparts. For highly-skewed log-normal distributions, however, the median* and mean* can be very far apart, depending on the range of possible values for the related measure. For example, in Figure 3.3 a distribution with SD* of 7.65 is shown with its median* of 27,274 and mean* of 216,280, which are both lower and much higher values than the standard median and mean (respectively) that would be derived by assuming a normal distribution. While these values may seem to over- and under-estimate the typical values that would be derived by traditional means, they more accurately reflect a dispersed range of typical values present in the log-normally distributed data.

Since all previous FM studies have assumed normal data distributions and thus reported normal means and SDs, we have facilitated comparison of our results with these studies by providing the normal mean and SD for log-normal data (excluding outliers) in addition to the log-normal statistics in each results table. This is especially useful for comparing measures exhibiting a large SD*, as discussed above. However, these figures are shown in parentheses to discourage their use because we advise caution in interpreting these values; though they enable some comparison to previous studies, for example when examining the growth of mean collection size over time, they can provide a misleading picture of the data when the SD* value is high.

Values of zero are removed in the derivation of log-normal statistics because the logarithm of zero is undefined. However, zeros may still be important to understanding user behaviour, for example in cases where a considerable portion of participants did not exhibit some behaviour at all, and so we report and discuss the number of zero values removed before the log-normal statistics were derived. In the discussion section we address the broader significance of having observed log-normal distributions in FM data.

Of our fifty-six measures of FM behaviour, forty-two are distributed log normally, eleven are distributed normally, and two have no obvious coherent distribution (they



(a) data histogram with normal (gray) and log-normal (red) fits, medians, and means

	mean	SD	median
normal	73,628	73,075	54,608
	mean*	SD*	median*
log normal	216,280	7.65	27,274

(b) normal and log-normal descriptions of the data visualised in (a), above

Figure 3.3 – An example of skewed, long-tailed data (a) and the resulting normal and log-normal descriptions (b). The high SD* shows a strong right skew, while the median* and mean* represent the lower and upper bounds of a *range* of typical values. This is more extreme in both directions (e.g., lower *and* higher typical values) than the standard mean and median suggest.

are long-tailed in virtue of being widely distributed but fit no distribution). The two measures' without typical distribution are discussed individually in the results, and in the tables only their zero values, maximums, and medians are reported. The flowchart in Figure 3.4 summarises the data preparation, classification, analysis, and reporting described here.

3.4 Results and discussion

By the end of our recruitment we had received 301 data files, all of which were usable, describing 42.9 million files across 6.6 million folders. Values resulting from analysing the data at 294 and 301 participants were, at most, 1% different, suggesting data saturation was reached. This sample included participants identifying as male (61%), female (37%), and other (2%), with ages ranging from 15 to 64 (mean 30, SD 9.8). Though we do not look for the potential effects of occupation on FM behaviour, our sample featured diverse occupations, including: poet, marketing director, electronics technician, bartender, data analyst, product designer, videographer, professor, safety inspector, and librarian. Data received came from laptops and desktops running Windows XP to 10, Mac OS 10.8 to 10.12, or one of eight Linux distributions, and were used for personal matters, work (or school), or both. The compositions of these and the demographic aspects of the sample are presented in Table 3.3.

	male	female	other
gender	185 (61%)	108 (37%)	5 (2%)
	range	mean	SD
age	15-64	30	9.8
	Windows	Mac OS	GNU/Linux
OS	135 (45%)	123 (41%)	43 (14%)
	laptop	desktop	other (tablet, server)
form	224 (74%)	74 (25%)	3 (1%)
	personal & work/school	work/school	personal
use	219 (73%)	50 (17%)	32 (11%)

Table 3.3 – Summary of population sample ($n = 301$) along demographic and technological characteristics.

The calculated results are presented below in three groups matching the standard

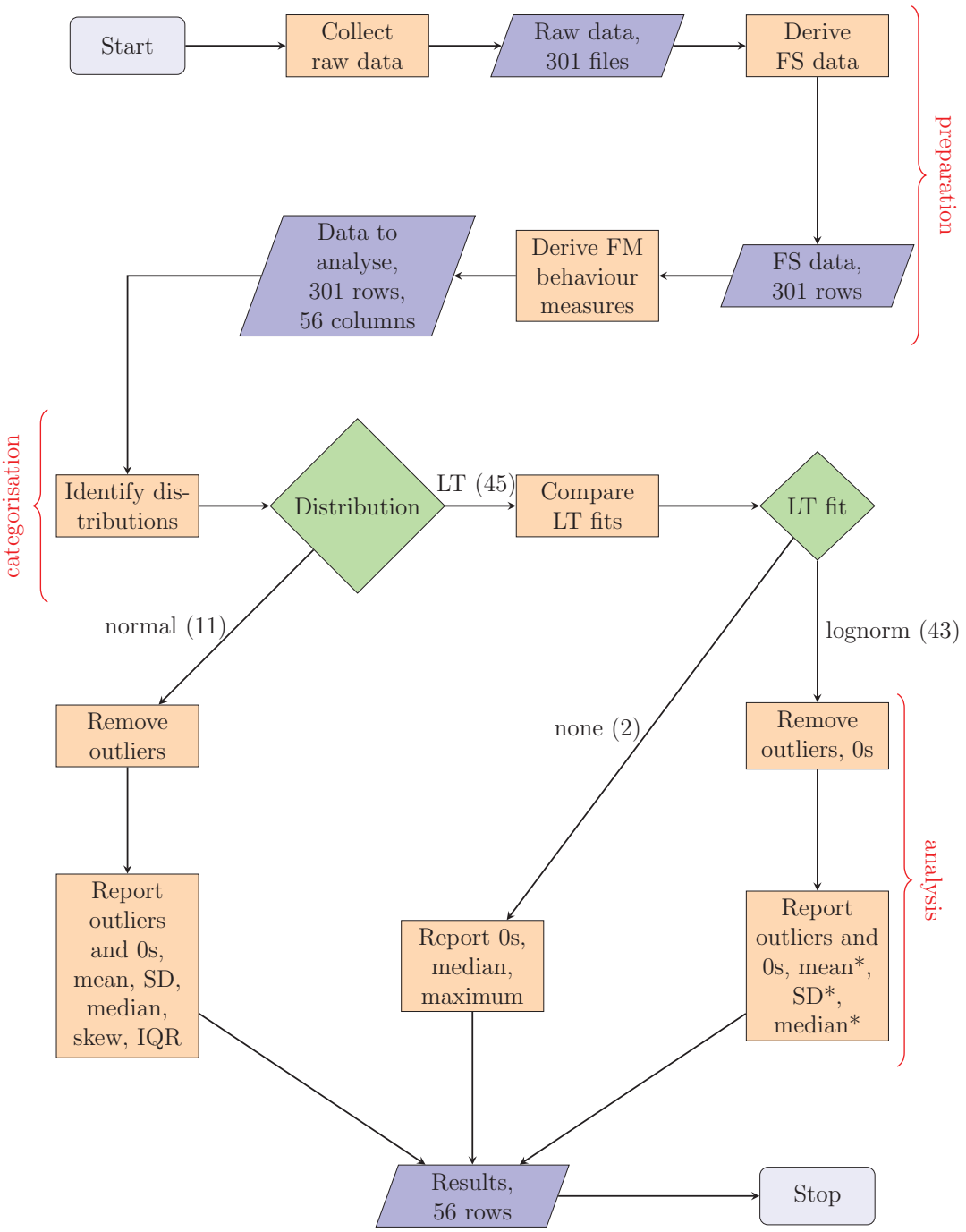


Figure 3.4 – Flowchart of data preparation, categorisation, and analysis, including relevant processes (rectangles), inputs and outputs (trapezoids), and decisions (diamonds). FS, FM, and LT refer to the file system properties, file management, and long-tailed distributions, respectively, and SD and IQR refer to standard deviation and inter-quartile range. Both visual and numerical analyses were used to identify distributions and compare fits.

PIM behaviour groups described in Table 3.8, comprising tabular data captioned with brief impression and followed in each subsection by a detailed discussion of the particular measures. Namely, storage behaviour is described in the tables showing measures

pertaining to participants' hardware, their digital collections' size and composition, and how they have named it; organisation behaviour is described in the tables showing measures pertaining to participants' folder structure and the categorisation of item within it; retrieval behaviour is described in one table showing its relevant measures. Within the tables, measures are displayed with the statistics appropriate to their distribution type, as discussed in the analysis above, and the traditional mean and SD have been added to the log-normally distributed measures to facilitate comparison to previous studies. Our results are therefore purely descriptive and no additional analyses have been performed; for example, we have not sought in this study to determine the effects of (or differences across) operating system, computer form, demographic features, or individual differences. Before presenting and discussing the results we address the significance of the majority of our data being log-normally distributed, as this affects how we compare our findings to those of previous works.

3.4.1 Log-normal distributions of data

Forty-three of our 56 measures, or 77%, were log-normally distributed, as discussed above, entailing a right-skewed trend in the data (i.e., the data trend towards the *left*) and typically a wide range of typical values and a long tail of high values. This is interesting for two reasons.

First, it implies that the data analyses of previous studies produced limited descriptions of the relevant phenomena and of *typical* behaviour. Many previous studies have assumed normal distributions of data, and described the data with a traditional mean and median despite observing very large SD values (e.g., Henderson and Srinivasan (2009); Hicks et al. (2008); Massey et al. (2014)), which suggested highly skewed data. By reporting traditional means rather than log-normal means and medians, these studies likely vastly underestimated the spread of central tendencies of their data and thus of typical behaviour. It is therefore advisable for future FM studies – and any PIM studies where varied behaviour is assumed – to check for long-tailed distribution and report the appropriate statistics. Doing so could more accurately describe the spread of typical values

and avoid the limited value given by the arithmetic mean.

Second, the wide spreads of typical values seen in log-normal data provide a quantitative confirmation of the frequent observation in PIM studies that PIM activity is highly personal and therefore varied (Lansdale, 1988; Jones et al., 2017), and suggests the same is true for file management. Nevertheless, people inevitably exhibit common behaviour during such activities and trends in this are reflected in the range of typical values described by log-normal measures. Such data can be used to quantitatively analyse the relative effects of the determining factors of FM behaviour, for example by first log-transforming the data[3] or adapting the data to negative binomial models (O'hara & Kotze, 2010). Before such relationships can be explored, however, an accurate description and understanding of the characteristics of the phenomenon in question is necessary, and such a description is more accurate when provided with log-normal means and medians.

3.4.2 Storage

Users' FM storage behaviour on a computer consists of keeping some number of files and folders across the available drives. Users may *actively* store files and folders by creating and downloading them, and *passively* by simply not deleting them, such as the folders that the operating system provides by default (e.g., Downloads, Pictures, My Music) and files that are unpacked from downloaded archives (e.g., zip and tar files). Regardless of how they are attained or kept, these occupy storage space on the user's device, be it a stationary or mobile device, provided by the available *drives* (e.g., hard disk or solid state drive). The capacity of these drives therefore provides a limit to the size of users' collections in terms of both the *number of items* and the *amount of disk space* they occupy, while the items stored have the potential to provide both utility – they contain or organise information – and frustration because they compete for the user's attention when a particular file or folder is sought. Here we present data relating to users' hardware, which describe their possibility to store data, and to their collections, which give an indication of the scale of the potential utility and frustration enabled by

[3]When this approach is taken in FM studies, for example in Massey et al. (2014), the descriptive values reported are still always the standard ones (e.g., mean rather than mean*).

their collection.

Hardware

measure	outliers	0s	mean*	SD*	median*	(mean)	(SD)
drives installed	5	0	1.93	1.7	1.68	(1.95)	(1.92)
drive capacity (GB)	31	0	632.05	2.61	399.16	(549.37)	(409.86)
space used (GB)	30	0	300.1	2.97	166.03	(246.88)	(206.29)
free space (GB)	25	0	399.2	4.1	147.45	(284.45)	(272.14)
collection size (GB)	31	0	257.96	7.78	31.44	(93.16)	(110.88)

Table 3.4 – Measures of participants' physical storage. Most participants have one or two drives totalling roughly 500 GB storage, with about half of that used. Space used and collection size differ because the former reflects space occupied by the OS and programs, while the latter reflects only the files and folders selected by participants for inclusion in the study. Data from every measure shown here were log-normally distributed.

The figures discussed in this subsection are presented in Table 3.4. Roughly 80% of the participants have only one or two hard drives, with nearly half having a single drive and just under a third having two. As 75% of our participants were using laptops, which often come with a single installed drive, this is not surprising, but suggests that desktop users typically have at least two hard drives installed. Regardless of form factor, participants typically have approximately 500 GB (or 0.5 TB) of possible storage space; given that most participants have one drive, this implies the presence of a standard 512 GB sized drive. These are roughly 25-50% full, and consequently, 50-75% of storage space remains, implying that users typically have sufficient storage space on their computers.

The available storage space we observed is much larger than that observed in previous studies, and this is not surprising given the constant increase in storage capacity available on hard drives from year to year since they became available. For example, in the 1990's fifteen knowledge workers had between 80 MB and 1.5 GB available (Nardi et al., 1995) and Microsoft employees had fewer than 2 GB (Douceur & Bolosky, 1999); employees at the same corporation a decade later had a median of only 40 GB (Agrawal et al., 2007). While drive capacity *can* limit storage, those same Microsoft drives were observed to be about half full (Douceur & Bolosky, 1999), regardless of job category, with a median of 42% full by the end of the studies (Agrawal et al., 2007). We observed similar use,

implying that users' ability to fill drives seems to be increasing with the drives' capacity. While participants in a previous study were adding drives to their computers as a result of running out of storage space (Barreau, 1995), this seems to be true only for desktop users today, as our results suggest that most laptops contain a single 512 GB that is less than half full.

In addition to the fullness of participants' hard drives we measured the stored size of their collections; while the former often includes the operating system and any other users' files stored on the same drive, the latter is the sum of all their files. Collection sizes vary even more greatly (SD* 7.78), typically ranging from 31 GB (median*) to 260 GB (mean*). This is far greater than the average size of engineers' collections, 2.5 GB, seen in the last decade (Hicks et al., 2008), but the upper bound of 260 GB suggests that most collections can currently be backed up onto a modestly-sized external hard drive.

File and folder storage

measure	outliers	0s	mean*	SD*	median*	(mean)	(SD)
collection size (items)	33	0	249,314	7.61	31,801	(86,004)	(85,390)
files	33	0	216,280	7.65	27,274	(73,628)	(73,076)
folders	35	0	26,181	7.62	3,331	(9,790)	(11,032)
file size (MB)	36	0	1.58	2.92	0.89	(1.39)	(1.25)
file age (days)	6	0	368.94	2.74	222.08	(332.37)	(262.43)
folder age (days)	7	0	354.73	2.7	216.54	(315)	(234.54)
hidden files	31	8	867.05	5.51	202.22	(463.18)	(501.17)
hidden folders	26	45	68.32	4.69	20.72	(39.37)	(47.53)
shortcuts	8	39	11,382	16.63	218.74	(1,502)	(2,277)
shortcuts (% of collection)	27	39	2%	7.43	0%	(1%)	(1%)
	0s	median	maximum				
hard links dupl. files	158	0	269,762				
hard links dupl. folders	136	316	399,432				

Table 3.5 – Measures of participants' storage of files and folders. Storage behaviour varies widely, but typically implies storing *at least* three thousand folders and nearly thirty thousand files. Shortcuts and hidden and hard links make up a very small portion of the collection, and most files are less than a year old and small in size. Data from most measures were log-normally distributed, while counts of hard links followed no coherent distribution.

The figures discussed in this subsection are presented in Table 3.5. While collections can be measured in terms of storage space, as above, typical FM interactions deal with

single or groups of files and folders. We measured the total number of files and total number of folders participants stored, and found that both were log-normally distributed and with similarly right-skewed shapes (SD* around 7.6).

Our figures are unsurprisingly greater than those reported in the 80's, where participants had 114 files on average (Carroll, 1982), though computer scientists had more, at 4.5 thousand (Satyanarayanan, 1981). This was seen to grow through the first decade of the 2000's to averages between 4.6 and 7.9 thousand (Gonçalves & Jorge, 2003b; Henderson & Srinivasan, 2009; Hicks et al., 2008), and recent PIM studies have seen larger figures still at 14 thousand (Whitham & Cruickshank, 2017, among academics), 15.3 thousand (Massey et al., 2014, among psychology students), and even 36.6 thousand files (Zhang & Hu, 2014, among information workers). With a traditional mean of 73.6 thousand, our data show a clear growth over time and a doubling in size from the previous largest figure in PIM literature.

The traditional mean does not reflect the typical case, however, as the log-normal statistics derived describe a wide range of common values. We found a median* of over 27 thousand and mean* of over 216 thousand, suggesting that the *typical* user has *at least* many more files than suggested by most previous studies, and possibly an order of magnitude more than suggested in any study.

Like files, folders must be stored, named, and organised, although they may also be navigated through and their internals sorted to provide access to files. Four PIM studies, all from the last decade, found varying means of the number of folders people kept, including 57 (Boardman & Sasse, 2004), 1 thousand (Henderson, 2005), 628 (Henderson & Srinivasan, 2009), and 2 thousand folders (Whitham & Cruickshank, 2017), while the study of Microsoft employees saw a much higher mean of 8.9 thousand (Agrawal et al., 2007). As with files, we observed a wide range (SD* 7.62) of typical values, but these were considerably larger than previously studies suggest: average FM behaviour currently consists of keeping at least 3.3 thousand folders (median*) – triple the largest number reported in previous PIM studies – and potentially an order of magnitude more at 26 thousand (mean*).

Unsurprisingly, similar figures are reflected in measuring participants' entire collections (e.g., files and folders together), and these collections are mostly made of files (about 86%).

Our participants also kept special items that have received little attention in previous studies, like shortcuts and hidden files. These made up a very small portion of their collections, however. For example, the typical percentage of participants' collections consisting of shortcuts is between 0 and 2% of the files; while the upper bound of this range is considerably more than the 0.06% reported in a previous study (Gonçalves & Jorge, 2003b), this is probably only noteworthy for exceptionally large collections since Bergman et al. (2010) found that folder shortcuts are used in only 11% of retrievals of working files. Roughly half of our participants stored hard links, but those that did stored very few of each (median 0 files, 316 folders). The lack of these items implies they are not a common part of FM, but also suggests further targeted research could reveal what is preventing users from making use of them.

The mean sizes and ages of items in participants' collection were also noted. Our participants' mean file sizes typically ranged from 890 KB (median*) to roughly one and a half MB (mean* 1.58), both of which are larger than reported in the next most recent studies, such as 189 KB among Microsoft employees (Agrawal et al., 2007) and 613 KB among a group of engineers (Hicks et al., 2008, SD 871), although these traditional means may have been underestimating the typical values of log-normally distributed data. Files were typically 222 (median*) to 369 (mean*) days old and folders were 6 to 14 days younger. This is consistent with the findings that most files have creation times in the last year (Gonçalves & Jorge, 2003b) while non-working files specifically are typically at least 6 months old (Ravasio et al., 2004), but we reiterate that such results should be interpreted cautiously since files' and folders' creation time metadata can be overwritten by applications or when the item is moved to a new drive.

measure	outliers	0s	mean*	SD*	median*	(mean)	(SD)
file name length	11	0	21.63	1.39	20.5	(21.58)	(6.72)
file name letters	19	0	12.12	1.35	11.6	(12.09)	(3.32)
file name numbers	3	0	6.7	2.41	4.55	(6.16)	(4.26)
file name specials	19	0	1.5	1.71	1.3	(1.46)	(0.65)
file name spaces	36	1	0.54	4.58	0.17	(0.33)	(0.32)
duplicate file names	30	6	127,296	11.33	6,686	(25,494)	(31,389)
folder name length	17	0	11.71	1.3	11.31	(11.7)	(3.06)
folder name letters	16	0	8.98	1.3	8.68	(8.97)	(2.30)
folder name numbers	17	5	1.81	2.26	1.3	(1.66)	(1.12)
folder name specials	18	6	0.59	1.83	0.5	(0.56)	(0.26)
folder name spaces	27	5	0.35	4.03	0.13	(0.25)	(0.27)
duplicate folder names	31	12	19,257	9.03	1,711	(5,480)	(6,640)
	outliers	0s	mean	SD	skew	IQR	median
% files duplicate names	1	6	29%	18%	0.59	25%	26%
% folders duplicate names	0	12	45%	21%	-0.41	27%	50%

Table 3.6 – Measures of participants' naming behaviour. File names tend to be 20-21 characters long and composed of letters, numbers, and special characters, and rarely feature a space despite their length. Folder names tend to be half as long and made mostly of letters. Duplicate folder names are very common, comprising nearly half of most collections, while duplicate file names make up nearly a third. Data from all but two measures here were log-normally distributed, while the percentages of all files and folders that had duplicate names were normally distributed.

File and folder naming

The figures discussed in this subsection are presented in Table 3.6. Another task in FM is managing the names of files and folders, for example by deriving descriptions of items, their content, or their functions when naming or renaming created or downloaded items. Users' approaches to this task can be studied by examining trends in the names' content (Bergman et al., 2006) or by making quantitative measures of users' naming behaviour (Carroll, 1982), which aids in understanding the goals motivating the behaviour.

We have taken the latter approach and found typical file names of 21 characters excluding the file extensions. File names were typically composed of 12 letters (57%), 5-7 numbers (24-33%), one or two special characters (e.g., punctuation or symbols), and sometimes a single space (mean* 0.54). This is greater than previously reported figures, and implies a growth of file names from mean length of 6 characters (Carroll, 1982, file names were limited to 8 characters at that time, and included file extensions) to 12.6 (Gonçalves & Jorge, 2003b) and 18.8 (Fitchett & Cockburn, 2015, included file

extensions). Our counts of letters, numbers, special characters, and spaces are roughly consistent with the observations of Gonçalves and Jorge (2003a), although they suggest a greater use of non-alphanumeric characters (e.g., the underscore).

To our knowledge no previous study has reported quantitative measures of folder names. Our participants' folder names were typically 11 or 12 characters long (roughly half the length of the typical file name), consisting of 9 letters (75%), 1 or 2 numbers, occasionally a special character (mean* 0.59), and rarely a space (mean* 0.35). This may reflect the tendency to name folders after the projects they represent or formats of the files stored within (Bergman et al., 2006; Jones et al., 2005), which are unlikely to contain many punctuation marks or numbers. This may also explain why folder names are shorter than file names: file names may be highly descriptive of their specific contents whereas folder names describe the project or format that unifies the files. Folder names also have a greater proportion of spaces than file names and a smaller proportion of special characters, implying that in file names spaces are replaced by punctuation like underscores or dashes.

We found counts of file and folder name duplication (defined as consisting of exactly the same characters) to be highly skewed, even more than counts of files and folders, with SD* of 11.33 and 9.03, respectively. Typical counts of duplicate file names ranged from 6.7 thousand (median*) to 127.3 thousand (mean*), and duplicate folder names ranged from 1.7 thousand to 19.2 thousand; compared with all files and folders, respectively, these duplicates composed 29% (SD 18%) and 45% (SD 21%) of their file and folder counts. Though the counts and proportions of duplicate names is higher than found by Henderson and Srinivasan (2009, 21.8% file name duplication), the proportion of duplicate file names is close to that found by Hicks et al. (2008, 32.4%). The proportion of duplicate folder names was considerably larger than those found in any previous study, such as 23.5% (Henderson & Srinivasan, 2009) and 31.3% (Hicks et al., 2008), with similar SDs, showing growth over time. While Henderson (2011) found that file and folder name duplications were correlated with deeper trees, but not with wider trees, it is currently unclear how collection size may determine this, but as file and folder counts

are increasing, the duplicated proportions suggest that file name uniqueness is stable throughout the growth while more folders are sharing names.

3.4.3 Organisation

We defined organisation behaviour as behaviour that involves structuring the folder tree or categorising files within it, roughly consistent with organising or meta-level behaviour (Jones, 2007b; Whittaker, 2011), and provide measures of each component. This consists of activities such as moving and arranging folders, which creates the outer shape of the folder tree, creating subfolders, moving and arranging subfolders and files within the tree, and leaving files and folders at the root. Users must then navigate through the resulting folder structure, and along the way encounter various combinations of files and folders that are products of their work.

Structure

measure	outliers	0s	mean*	SD*	median*	(mean)	(SD)
roots	14	0	1.47	1.51	1.35	(1.48)	(0.69)
max tree breadth	39	0	4,439	6.31	813	(2,115)	(2,303)
mean breadth	31	1	853	4.6	266	(568)	(589)
folders at root	22	1	18.3	1.85	15.16	(17.85)	(10.02)
leaf folders	37	0	18,730	7.73	2,315	(6,829)	(7,648)
branching factor	17	0	3.62	1.28	3.51	(3.61)	(0.85)
switch folders	23	5	5,980	8.34	630	(2,113)	(2650)
	outliers	0s	mean	SD	skew	IQR	median
max tree depth	1	1	15.52	6.06	-0.37	9	16.5
waist depth	1	3	6.16	2.67	0.27	4	6
mean folder depth	5	0	6.61	2.01	-0.63	2.48	7.09
mean leaf folder depth	6	0	6.81	1.98	-0.61	2.37	7.31
% of leaves	12	0	72%	6%	0.15	9%	72%
mean switch folder depth	15	0	6.53	1.85	-0.45	2.16	6.74
% of switch folders	5	5	16%	7%	-0.01	11%	16%

Table 3.7 – Measures of participants' structuring behaviour. Folder trees start at a root and typically immediately spread to 15-18 branches, then spread at a rate of 3.5 branches per subfolder, extending to be hundreds of folders wide and 15 folders deep. Most folders are leaves located 6-7 levels down, also the widest part of the tree, implying a roughly lozenge shape. Data from half of the measures are log-normally distributed, while the other half, mostly measures of depth, are distributed normally.

The figures discussed in this subsection are presented in Table 3.7. Most of our participants' collections' had one or two *roots* (mean* 1.47): roughly 58% had only one root, 25% had two, and 10% had three. These numbers align with the number of hard drives our participants had, discussed above, suggesting that users' collections are typically in one or two folder trees stored on as many drives, rather than multiple trees stored on one drive. This is roughly consistent with the numbers of roots seen in previous studies, for example 1.19 observed by Hicks et al. (2008), and with the finding that an information *locus*, such as a single device, typically has one high level folder (Gonçalves & Jorge, 2003b). This differs from the mean of 3.4 roots reported by Henderson and Srinivasan (2009), but this is likely due to the definition of *root* used in that study, which included participants' desktops and My Documents folders.

Our participants' folder trees exhibited a mean maximum depth (or height) of 15.52 (SD 6.06; excluding only 1 outlier), almost twice deeper (or taller) than the greatest of the previously reported figures, which range from 4.0 to 8.67 (Gonçalves & Jorge, 2003b; Hicks et al., 2008; Henderson & Srinivasan, 2009; Henderson, 2011; Zhang & Hu, 2014). At the root level or top of the tree we found a typical range of 15 to 18 folders, suggesting the top of the tree, the most traversed part, is somewhat wide and descends into several subtrees. This is roughly consistent with the previously reported average of 19 (Khoo et al., 2007), while higher figures are reported in studies regarding the desktop and My Documents folder as roots (Henderson & Srinivasan, 2011).

Moving down in the tree, we saw that the typical maximum tree breadth (or width) varied greatly, with typical values ranging from 813 (median*) to 4.4 thousand (mean*), and the mean breadth at any given depth ranged from 266 folders (median*) to 853 (mean*). This entails that trees are relatively wide, and much wider than deep, contradicting a previous conclusion that trees are narrow (Gonçalves & Jorge, 2003b). The mean depth of the widest point was 6.16, or just past one third of the way down the tree from the root, and accordingly the mean depth of folders in the tree was nearby, at 6.61. This is roughly consistent with the two previously reported mean folder depths of 5.12 (Zhang & Hu, 2014) and 6.9 (Agrawal et al., 2007), but considerably deeper than

previous figures of 3.3 (Boardman & Sasse, 2004) and 2.5 (Zhang & Hu, 2014, two groups of participants were analysed in this study).

Looking to the bottom of the tree, we found a wide range of typical counts of leaf folders, from 2,315 to 18,730, which is likely attributable to the widely varying number of folders participants kept (median* 3,331 and mean* 26,181, as seen in Table 3.5). Per each participant's total folder collection, leaves accounted for a mean of 72% of folders (SD 6%), which is close to figures ranging from 65 to 70% previously reported (Douceur & Bolosky, 1999; Agrawal et al., 2007). We observed that the mean depth of leaf folders was near to the waist (6.81), implying that the bottom of the tree starts just below the waist and much of the bottom exist there, such that the tree must taper in width towards a point at the deepest depth. As a previous study has reported relatively normally-distributed frequencies of folders across the depths of the folder tree (Hicks et al., 2008), this implies most trees have a shape approximating a diamond (or lozenge).

The internal structure of the tree is described by a measure called *branching factor*, defined as the mean subfolders per non-leaf folder. This gives an indication of the average complexity of a navigation decision made within the tree; for example, a branching factor of 2 would mean the tree typically *branches* in two directions, making the typical decision for a user navigating the tree to be between two possible paths downwards in the tree's structure. We found typical branching factors range from 3.51 (median*) to 3.62 (mean*), implying the typical navigation decision entails choosing between three or four folders. This is consistent with previous figures from comparable contexts, which reported branching factors of 4.0 (Henderson & Srinivasan, 2009) and 3.4 (Zhang & Hu, 2014). This is considerably lower than reported branching factors describing the first few depths of trees, for example 8.13 among trees that were on average only 4 levels deep (Zhang & Hu, 2014) and 10.64 observed during navigation starting from the top of tree (Bergman et al., 2010), which suggests that the first few levels of a tree branch quickly, consistent with the diamond shape described above but not with the idea of a narrow tree identified previously. As we report above that the root contains 15-18 folders, this also means that the first navigation decision made when starting from the root involves

at least 10 folders more than a decision made at any other given folder below the first few levels.

Some folders are used primarily to guide navigation and house other folders (Akin et al., 1987; Bergman et al., 2010). We defined these strictly as folders containing any number of subfolders but no files, and found that their counts varied greatly, as the counts of folders did, but that they composed an average of 16% of participants' folder trees (SD 7%). This shows an increase in the occurrence of switch folders from roughly 6% of the tree in their first measurement (Akin et al., 1987). The mean depth of such folders is 6.53 (SD 1.85), quite close to the depth where the tree is widest (i.e., waist depth). The measures of folder depth for switch folders, leaf folders, and all folders together were normally distributed with slight negative skews and similar SDs, implying folders are found at roughly the same frequency throughout the folder tree regardless of being leaves or for navigation.

Categorisation

measure	outliers	0s	mean*	SD*	median*	(mean)	(SD)
files at file waist	28	0	56,970	6.67	9,416	(23,959)	(24,168)
files per folder	22	0	7.96	1.68	6.96	(7.9)	(3.95)
empty folders	23	11	3,683	9.18	315	(1,092)	(1,237)
% empty folders	23	11	13%	3.16	7%	(10%)	(8%)
unfiled files	42	79	8.5	2.95	4.73	(5.66)	(7.98)
root pile rate	56	79	0.002%	4.75	0%	(0%)	(0%)
	outliers	0s	mean	SD	skew	IQR	median
depth of files	9	0	6.1	1.77	-0.49	2.19	6.3
file waist depth	5	2	5.6	2.29	-0.04	3	6

Table 3.8 – Measures of participants' categorisation behaviour. Most files are filed into folders five or six levels deep into the tree, with very few (5 to 9) being located at the root. 7 to 13% of folders are empty, while most contain 7 or 8 files. Data along most measures are log-normally distributed, while two measures, both of depth, are distributed normally.

The figures discussed in this subsection are presented in Table 3.8. Physical, paper documents may be organised into files or left *piled* on the desk (Malone, 1983). This distinction, of filing and piling, and the corresponding distinction of people who exhibit such behaviour as filers or pilers, has extended to the digital world together with other

categorisation strategies that have been the subject of several previous studies focusing on contexts including email, Web bookmarks, and digital files (Oh, 2017 provides a broad overview). Digital files in particular are categorised by being placed by the user into various locations in the tree, or are left uncategorised (also *unfiled*), which by definition means being left (piled) at a *root*, such as the user's home folder or root of a drive.

We found that 79 participants (26%) had no unfiled files, and those that did typically left fewer than 10 files unfiled (median* 4.7, mean* 8.5), producing a root *pile rate* (percentage of unfiled files to all files) of 0.002% or lower. This is lower than the 2 to 3% previously seen in FM contexts (Boardman & Sasse, 2004; Henderson, 2011; Henderson & Srinivasan, 2011). This is likely because previous studies included files on the desktop (Boardman & Sasse, 2004) and in the My Documents folder, whereas our software treated these as roots only if users specified them as such (described in detail in the methodology, above). Regardless, their figures and ours are both low, suggesting that while piling information is common in digital contexts such as emails and Web bookmarks (Boardman & Sasse, 2004) or online learning environments (Hardof-Jaffe et al., 2009a), it is not a common behaviour in FM. This is likely due to differences in the collections: users may be more invested in organising files because they encourage a stronger sense of ownership (Boardman & Sasse, 2004) and may adapt their organising strategy to the foreseen retrieval task of the relevant items (e.g., searching for Web pages rather than navigating to files). Definitions of piling that are more inclusive (e.g,. regarding piling as placing files in places beyond the root) and further analyses of files *per* depth may reveal more nuanced results, but it is possible that the very presence of so many files is *requiring* filing simply to maintain the collections' comprehension, accessibility, and navigability. This seems especially likely given the high redundancy of file and folder names, which reduces the utility of searching for files by name and thus encourages location-based categorisation into files.

We found that the mean depth of all files was 6.1 (SD 1.77), exactly the average tree waist depth, with the typical number of files at that depth ranging from 9.4 thousand (median*) to 57 thousand (mean*). However, we also found that the *files waist*, defined

by the mode of file depths, was at 5.6 (SD 2.29), suggesting the widest part of the tree in terms of files is slightly above the widest part of the tree in terms of folders. Perhaps because of our Windows participants (45.9%), we observed slight peaks at depths of 2, 3, and 5 consistent with studies of exclusively Windows participants (Agrawal et al., 2007; Hicks et al., 2008), but interestingly did not find the reported peaks at depths of 4 nor 7. Studies observing only recently accessed files found lower mean depths, ranging from 1.81 to 3.7 (Bergman et al., 2010; Bergman, Whittaker, & Falk, 2014; Fitchett & Cockburn, 2015), suggesting that files at or below the tree waist or file waist are likely accessed less frequently. Storing less-frequently accessed files deeper in the tree may be an adaptive behaviour to aid refinding them by using the descriptive reminders provided by folders about the content they contain (Barreau & Nardi, 1995).

Folder fullness, or the typical number of files per folder, provides a measure of difficulty for the average target-identification task in FM. We found the typical number of files per folder to range from 7 to 8 (SD* 1.68; traditional mean 7.9 and SD 3.95), slightly fewer than reported in previous studies from 1987 to 2014, which tend to report 11 to 12 files per folder (Akin et al., 1987; Hicks et al., 2008; Henderson & Srinivasan, 2009; Bergman et al., 2010; Zhang & Hu, 2014), while two have reported 13 (Gonçalves & Jorge, 2003b) and 18.9 (Massey et al., 2014). As our data produce similar figures if we include 22 outliers and assume a normal distribution, producing a mean of 10.09 (SD 10.93), it may be that previous studies were assuming normality of log-normal data and including extreme outliers (c.f., Massey et al. (2014), where the mean of 18.9 was accompanied by a SD of 16.4). Regardless, our findings of 7 to 8 files and 3 to 4 folders per folder imply 10 to 12 items per folder, consistent with the mean of 11.82 found by Bergman et al. (2010). This suggests a stability in categorisation over the last decade and entails that users continue to keep fewer items per folder than Bergman et al. (2010) found would incur retrieval problems (i.e., fewer than 21 items).

Folders may also be *empty*, containing no files or folders. These may be made, for example, by putting nothing in them at the point of their creation, perhaps in anticipation of forthcoming projects (Khoo et al., 2007), or by not deleting them when the last file

or folder is removed. While 11 of our participants (3.6%) had no empty folders, most did, with the range of typical counts varying slightly more than the total number of all folders (SD* 9.18), and typically comprising 7% to 13% of the entire folder tree (SD* 3.16). This is near to previous figures of 8% (Henderson & Srinivasan, 2009) and 18% (Douceur & Bolosky, 1999), suggesting stability of the proportion of empty folders over time and collection growth, which is surprising given that Henderson and Srinivasan (2009) found that collection size had no effect on the number of empty folders. Further work is required to confirm the origin of such folders, clarify their relation to collection size, and understand the impact they have on regular FM tasks.

3.4.4 Retrieval

measure	outliers	0s	mean*	SD*	median*	(mean)	(SD)
time since file access (days)	21	0	271	3.64	118	(202)	(173)
time since file modify (days)	17	0	754	2.99	414	(609)	(438)
time since folder modify (days)	15	0	413	3.23	208	(323)	(244)

Table 3.9 – Measures of participants' retrieval behaviour. File access behaviour varies widely, with a range of typical values indicating most files have not been accessed for at least four months or modified for over a year, while most folders have been modified more recently. Data from all three measures here are log-normally distributed.

The figures discussed in this subsection are presented in Table 3.9. Few traces of users' behaviour are left after they retrieve files, but metadata associated with each file and folder tells the last time it was accessed or modified. As with file creation times, the values in such properties should be interpreted cautiously as they can be overwritten by software without the user's initiation or knowledge (this is true of all operating systems and consequently of the data in all similar studies). We do not have data about when the folders were last accessed because when our software entered each folder to examine its contents, that property was updated to the current time by the operating system (i.e., our program was accessing the folders).

Files were typically last accessed between 118 and 271 days (or roughly 4 to 9 months) prior to our data collection and modified even longer ago, typically 414 to 754 days (1.12

to 2 years). Folders were typically modified more recently, ranging from 208 to 413 days (7 to 14 months). These figures describe behaviour roughly consistent with the previous observation that while most folders have not been modified within the last month, most files have been accessed (but not modified) in the past year (Gonçalves & Jorge, 2003b). It also suggests more frequent file access than that observed in engineers (Hicks et al., 2008), who seemed to have not accessed the majority of their files in the last year.

While these measures provide a narrow depiction of retrieval behaviour, they nonetheless suggest that the majority of any given collection has not been accessed recently. Recently accessed files have been the main targets of retrieval tasks given to participants in observational PIM studies (Bergman et al., 2010, 2012; Bergman, Whittaker, & Falk, 2014; Fitchett & Cockburn, 2015, for example). While these files deserve the scholarly attention they have received, the results of such studies may not apply to the majority of participants' collections.

3.4.5 Implications

The goal of our study was to provide a broad quantitative description of typical FM behaviour. With the tabular data above we have described such behaviour along 56 measures, and three general findings about typical behaviour became evident in our discussion. First, users' are storing considerably more folders and files (an order of magnitude more) in broader and longer trees than previously reported. This does not seem to have an effect on the general consumption of storage space, but may entail new management challenges due to the sheer size of the collections; even a search for known file name may return too many files to easily review, especially given the immense duplication of file and folder names. This also generally implies that users are in fact *doing file management*; even if most of a participant's files and folders were to come from pre-downloaded packages, that would indicate that they are, at the very least, choosing to acquire and store these collections. The more likely explanation is that users are, over time, creating and managing these collections and structures, and the sheer number of files and folders stored implies that despite the advances in desktop search and tagging

and the perception that files are old fashioned, people are still storing, organising, and retrieving files and folders.

Second, despite the growing collections observed here, the internal folder tree structure and file categorisation have remained consistent. The cause of this is unclear, but investigations into users' cognitive dispositions for certain organisational strategies (Oh, 2017, for example) may provide further insight, as would targeted qualitative studies examining users' tendency to keep fewer than twenty-one items per folder and to create so many empty folders. While some of the differences in the organisation strategies users employ across various contexts (e.g., filing or piling their files, emails, or Web bookmarks) have been attributed to the contexts themselves (Boardman & Sasse, 2004), the exact causes of the stable organisation and prevalence of filing observed here should be studied further as it may reveal users' preferences for knowledge to be organised similarly in other information access, browsing, and retrieval contexts, for example in online repositories or subject heading trees (e.g., LCSH or MeSH).

Third, file access behaviour seems stable, with the average file accessed 4-9 months ago but not modified in the last year. This implies that recently accessed files do not represent users' collections, and so the non-working or archived portions such collections should be examined more closely in future studies. Finally, we found narrower but noteworthy observations: despite the functionality offered by special items like shortcuts and hard links, they make up a negligible proportion of participants' collections, implying little use. Understanding what prevents people from using such items would benefit from targeted qualitative study.

These findings and the data reported in our tables together describe typical FM behaviour, thus constituting an answer to RQ1, and the differences in the findings identified against previous studies as seen across various measures, discussed above, constitute an answer to RQ2. Our results and their discussion provide a common point of comparison across the disparate quantitative descriptions given in previous studies, and can complement the rich qualitative descriptions of FM given in previous PIM studies to form a more complete picture of FM. The results therefore facilitate advanced methods for

studying FM such as principal component analysis, user modelling, and investigating the relative strength of factors determining FM behaviour. Identifying the possible measures and reporting their respective data, as we have done here, is a necessary step towards such work.

By comparing the results of future studies of particular aspects of FM behaviour to the data reported here, changes in behaviour can be tracked over time to understand, for example, the effects caused by changes in file size and richness, the commonality of multimedia files, the availability of greater storage capacity, the existing and new features and restrictions of operating systems and file management software, the demands and conventions of various occupations, and the individual differences exhibited by computer users. The data reported here may also aid in the interpretation of observed and self-reported user behaviour, for example by determining if it is representative, and solicited user opinions, which may not reflect their behaviour (Bergman, Gradovitch, et al., 2013b). For example, given the average file size (1.6 MB or smaller) and collection size (30 to 260 GB) seen here, cases of study participants reporting that files are sometimes too big to easily back up (Kljun, Mariani, & Dix, 2015a) or transfer between computers (Capra, 2009) may regarded as outlying cases.

Many measures (54) were used in this study to characterise common PIM behaviour groups (i.e., storing, organising, and retrieving); while future studies of PIM may not be concerned with every measure, by virtue of their breadth these measures and the resulting characterisations are likely to provide points of comparison for the results of PIM studies set in contexts beyond file management. For example, it may be useful to compare the structure of folders within Web-based email clients for storing and managing email to the use of traditional desktop folders, or to understand the scale of Web bookmark management by comparing it to the file storage measures reported here.

Our identification, categorisation, and use of many existing and new measures in this study aimed to provide a broad picture of FM behaviour and unify previous studies of FM that reported such measures piecemeal as they were needed to answer particular research questions. The availability of this large group of measures invites future work to

determine the relative usefulness and saliency of individual measures for characterising respective behaviour groups.

3.5 Limitations

We acknowledge various limitations to our approach and findings. Notably, by virtue of the quantitative approach used our findings consist of inferences about user behaviour drawn from measures of file system metadata that lack the context of user opinions and reports, direct observations of user behaviour, and knowledge of their particular personal, occupational, and computing contexts. This entails that our interpretations of the data collected, while informed by the findings of previous studies when possible, nonetheless lack the interpretive support provided by qualitative descriptions of user behaviour, such as users' explanations of specific instances of folder use (e.g., as was done by Whitham and Cruickshank (2017) and in many previous studies). As such, our interpretations of the results involve a large degree of subjectivity and should be considered cautiously. Such interpretations should, for example, be triangulated with complementary observations of user behaviour before being used to inform the design of information management systems or policies.

Our data were gathered in single snapshots of users' collections, and so give no indication of actions like deleting, renaming, or sharing files, and provide only a brief description of retrieval behaviour. By contrast, studies observing users as they retrieve files can note if files were navigated to or searched, how long the retrievals took, and if they were successful or not.

The analysis provided here is purely descriptive, and thus gives little indication of what portion of the data is caused by various internal and external factors, including OS, demographics and occupation, and psychological factors, and does not describe an analysis of the file types seen or the use of default folders observed. Now that a broad quantitative description of general FM behaviour has been established, it can be analysed to understand the relative impact of various internal and external factors. Fortunately,

we collected data about several such phenomena, including demographic, psychological, and technological traits, and so their analyses will follow in future studies. The data analysis method used in this study was also intended to describe typical cases, and thus we ruled out outliers. While outliers were typically uncommon, they likely still provide very interesting insights into user behaviour, and so deserve further analysis and attention in future studies.

Finally, a notable limitation is that the Library folder present in the home directory of participants using Mac OS (i.e., with the path /Users/<user>/Library) was included in the data collection. The Library folder is a hidden folder where applications store, arrange, name, and access files related to each user's profile in that application, and it is unclear if users are accessing or managing files within this folder. Although hidden, this folder is hidden with a Mac OS-specific flag, rather than having a name with a leading dot (as is the POSIX standard for hidden files), which was the only method by which the data collection software identified and ignored hidden folders at the time the data for this study were collected[4]. Consequently, the Library folder may have been included in the data collection for participants running Mac OS, and so a portion of the data describing those participants' collections (n=123, or 41% of the sample) is likely managed not by those participants but by their applications (i.e., app-managed rather than user-managed files).

It is unclear what portion of the data collected from participants using Mac consists of such files and what is the relative effect of the Library folder on our overall measurements and findings. In a preliminary analysis of a subset of the data excluding Mac participants' data, we found that all identified distributions (e.g., log normal or otherwise) did not change, but that the resulting values for FM behaviour measures changed in varying degrees across the measures; however, it cannot be conclusively determined from the present data how much of that change is caused by the Library folder or, as suggested in previous studies, by the effect of the OS on users' behaviour. Thus, the data reported and findings and implications discussed in this study should be regarded cautiously, and

[4]Cardinal (i.e., the data collection software developed and used in this thesis) has since been updated to exclude the Library folder from data collection.

future work should focus on generating a quantitative description of Mac users' FM behaviour with increased accuracy in excluding app-managed folders, especially the Mac OS Library folder.

3.6 Conclusion

File management is one of the most common cases of information management, and understanding and supporting it benefits from qualitative and quantitative descriptions of user behaviour, the latter of which we have provided here. This is an important step towards, for example, modelling FM behaviour and developing a standardised collection for FM system evaluation. Such methods are necessary to increase our understanding of users and develop advanced systems that are effective at supporting them in managing and browsing large digital collections. By successfully using our novel data collection tool, Cardinal, we have also demonstrated its potential for studying FM and for describing users' behaviour.

We found that FM behaviour varies greatly, as was evidenced by log-normally distributed data; this reflects the more general phenomenon that PIM behaviour is highly personal, and data about such behaviour requires special analyses to accurately describe and understand. While large standard deviations reported in previous comparable studies suggest that past data were similarly distributed, statistical analyses in such studies treated their data as normally distributed and thus underestimated the typical values and their spread. Future studies should therefore adopt the analyses demonstrated here to accurately describe log-normal data, and transform their data as needed for parametric testing.

We saw that while collections are growing over time, their internal folder structure and file categorisation remains stable. The cause and limits of this phenomenon should be explored, as should its implications for other cases of large, structured information sets. Now that a baseline of typical FM behaviour has been established, the relevant determinant internal and external factors should be explored. Following this lead, we

will analyse our data to explore the potential effects of the operating system, hardware factors like computer format and available drive space, and individual differences like occupation, spatial ability, and personality style.

Further work is required to fully understand the implications of our findings for the design of FM software and beyond to related topics of study. File management software was first designed when people kept very few files and folders on shared workstations and managed them with textual commands, and its modern, graphical counterparts, such as Mac's Finder program, have offered a stable core of functionalities since their early versions from the 1980's despite collection sizes growing exponentially. As our knowledge of FM behaviour becomes more detailed and nuanced, so can the support for it offered by such software.

In summary, we found that FM entails the highly varied behaviour assumed of most PIM contexts, and that improved desktop search, support for tagging, and the perception that files are old fashioned or obsolete have not caused users to give up the traditional management of files into folders. This supports the argument made in Chapter 1: as a collection grows in size, categorisation becomes necessary to keep it easily comprehensible, navigable, and accessible, and folders provide this fundamental need categorising the many files people are storing. Following this, it may be useful to investigate how and why behaviour differs in contexts that resemble or even take place within file management, such as the management of emails, Web bookmarks, and digital documents, photos, music, and references, and to see how certain personal digital archiving practices can be understood in relation to everyday file management. Studies of knowledge organisation and information behaviour have shown concern with the usability of, for example, large tree structures like the LCSH (Julien et al., 2013). FM constitutes a common case of information management and seeking, wherein users evidently create massive structures for storing, accessing, and understanding their digital collections, and we now know that users are accustomed to regularly navigating and retrieving from their own large structures. While managing collections of tens or hundreds of thousands of digital items, and classifying the items into and retrieving them from several thousand folders (subjects,

nodes, etc.) that were arranged by the user may sound like extreme cases of information management, our results suggest computer users are in fact doing this.

Conclusion

The goal of this thesis was to understand and improve upon the state of knowledge about file management (FM). This was done by meeting particular objectives, enumerated in the introduction and presented in Table 3.10 with the chapters in which the objectives are met and their respective theoretical and practical contributions. The remainder of this conclusion summarises the results, limitations, and implications of the thesis.

	Thesis objective	Contributions made	Chapter
1.	identify and synthesise studies of file management (FM), including identifying their common motivations and methods	*theoretical contributions*: increased understanding (e.g., identification, demarcation, characterisation) of field of FM research; synthesis and summary of motivations and methods of over 200 works	Ch. 1
2.	describe the state of knowledge among relevant studies, including identifying their findings, limitations, gaps in knowledge, and future directions	*theoretical contributions*: synthesis and summary of knowledge (e.g., findings, limitations, gaps in knowledge) resulting from FM studies; identification of necessary future research directions and tools to enable such research	Ch. 1
3.	develop software necessary to treat knowledge gap and alleviate limitations of quantitative data collection tools used in previous FM studies	*practical contributions*: design, creation, testing, and sharing of improved, extensible, and reusable data collection software that overcomes identified limitations; *theoretical contributions*: classification of file and folder metadata into established personal information management (PIM) behaviour categories (i.e., storage, organisation, and retrieval)	Ch. 2
4.	provide an extensive quantitative description of typical FM behaviour to enable further research like user and collection modelling and development of relevant theory, complement existing qualitative knowledge, and facilitate designing improvements to relevant software	*theoretical contributions*: extensive description of typical FM behaviour along many actions related to storage, organisation, and retrieval of information; characterisations of storage, organisation, and retrieval behaviour (notably: growing storage, stable organisation, frequent retrieval); identification of trends in user behaviour across previously disparate studies; *practical contributions*: demonstration of use of data collection tool; identification of need for and demonstration of analyses appropriate for describing log-normally distributed data describing FM behaviour; identification of trends in user behaviour that warrant future study and may benefit from targeted software design)	Ch. 3

Table 3.10 – The objectives of this thesis (all were met), with the resulting contributions made and the chapters in which these are found

Beginning with an extensive review of research related to FM (Chapter 1), the work

presented in this thesis identifies over 200 publications relevant to FM and reports their common research motivations and methods. The field of FM research is therefore now demarcated, and its main motivations identified as understanding how users do FM, what determines their behaviour, and attempting to aid them through novel interfaces and services (thus achieving thesis objective 1). The common methods of FM research are reported to include asking users about their behaviour and challenges, observing their behaviour directly, and inferring their behaviour from file system properties. The limitations imposed on inferring behaviour by existing quantitative data collection tools are identified, namely, allowing only narrow data collection (e.g., requiring small samples and being restricted to a narrow range of data) and being impractical to administer (thus achieving thesis objective 2). The results of these limitations are identified, composing various gaps in knowledge, including a broad quantitative description of users' FM behaviour that is necessary for various avenues of further study. Necessary and promising future directions for research were identified, including modelling users and their collections, and investigating the effects of factors like individual differences on FM, such as personality style, spatial ability, cognitive style. Also identified were the improvements required to data collection tools to enable such future studies and improve the overall knowledge about users' FM behaviour.

Novel data collection software was then developed specifically to overcome the limitations identified in existing tools (Chapter 2; thus achieving thesis objective 3). Specifically, its design enabled recording a large number of file and folder metadata – making 38 of 40 conceivable measures relevant to and categorised into established personal information management (PIM) behaviour categories – and facilitated large-scale and wide-spread data collection through the distribution of cross-platform binaries offering participants remote, asynchronous, and anonymous participation. The software's efficacy, efficiency, and improvement over existing data collections tools were demonstrated in a pilot study, and its source code was then shared with the research community to enable modifying the tool and using it in future studies.

That data collection software was then employed in a study of the FM behaviour of

301 people (Chapter 3), resulting in a broad quantitative description of FM behaviour composed of 56 measures of behaviour (as compared with the previously largest 14 measures), thus enabling future research (and achieving thesis objective 4). To provide this description, the data analyses required for a meaningful description of FM behaviour data were identified and employed, resulting in a demonstration of the identification of data distributions and the use of analyses appropriate to log-normally distributed data. This was the first time such analyses were used in PIM research, despite similar data being reported previously, and so the implications of this for future PIM research were explicated: the highly-varied behaviour seen in PIM often results in highly-skewed data, and in such cases, analyses like the ones demonstrated should be done to provide a sound description and avoid underestimating the values associated with such behaviour and its range.

The behaviour observed was characterised and compared to previous studies, with the primary findings being, in summary: 1. despite the availability of alternative modes of interacting with digital content that are believed by the research community to soon replace files and folders (e.g., tagging, search), people are keeping many files and organising these into many folders, and are doing so considerably more than what has been observed previously, 2. while the folder structure they use to organise these items is growing, its internal properties (e.g., file categorisation and number of subfolders per folder) remain the same, 3. most files and folders have not been accessed or modified in the last six months, but most have been accessed in the last year, implying both (a) that since people keep a large number of files, they must perform a considerable number of retrievals in a year, and (b) that recently accessed files are not representative of participants' collections.

In summary, the results of this work are a characterisation of a large but previously unidentified body of research into a ubiquitous information management activity (i.e., FM), analyses of the findings, importance, and necessary future directions of that field, the development of a tool necessary to advance research in some of those directions, and the first of many steps in those directions.

The work presented here has notable limitations, however. Perhaps most obviously,

while acknowledging the role of qualitative aspects of FM behaviour, this work narrowly operationalises behaviour in purely quantitative measures derived from records of file and folder metadata that serve as an artefact of user behaviour. This necessarily excludes many additional aspects of FM that are worthy of study, sometimes called *contextual factors*, which may include, for example, the occupational demands, user beliefs, or properties of the information stored that potentially determine why users store or manage their content in the way they do. Our approach also necessarily limits our conclusions to what is observable and can be inferred from the file system, and only as the file system was seen at the moment data collection was performed. This entails that we cannot examine users' actions that leave no trace in the file system, such as deleting files, that we cannot discern the difference between actions that leave the same evidence, such as those that change a file's modify time (e.g., moving or renaming a file), and that we cannot account for actions taken across multiple devices. The snapshot of the file system and thus of a participant's FM behaviour do not directly inform us of, for example, the evolution of files' names, locations, or sizes over time.

The conception of FM behaviour adopted for this thesis, including the operationalisation of FM behaviour and the quantitative data collected, was chosen to treat the perceived limitations of previous FM studies, discussed above. For example, it is only by using an automated, quantitative method that we were able to collect data about millions of files from hundreds of participants, which is needed to form a basic description of typical FM behaviour (as we defined it). It is our hope that our approach complements the qualitative ones currently in use in FM research, and that our findings in interpreting their results. Through a triangulation of qualitative and quantitative investigation, for example by asking users why they do FM in the way they do, and interpreting their responses with the help of accurate measurements of their behaviour.

Another notable limitation is the possible inclusion of application-managed files (e.g., those in the Library folder) in the data collected from participants using Mac OS. This limitation, discussed in Chapter 3, entails that our findings should be interpreted cautiously and verified in future work.

Despite its limitations, but in virtue of its research design, our work presents a broad quantitative description of what constitutes typical FM behaviour. This enables further work, such as modelling users and their collections and identifying the principal components of FM behaviour, which are necessary to increase our understanding of and ability to support the daily task of FM. With the *what* of FM now established, future research should examine the *why* by identifying and understanding the relative effects of internal and external influencing factors. This includes, as discussed in Chapters 1 and 3, individual differences like demographic features and psychological traits, contextual factors like occupational demands and collaborative information management, and external factors like the hardware and software used to perform FM. As supporting users is the end goal of FM research, future studies should also explicate and expand on the possible explicit software design guidelines that can be inferred from the detailed description offered here.

This thesis, through its findings and methodological contributions, constitutes an improvement upon the existing knowledge about FM, and facilitates further improving that knowledge. As our understanding of FM increases, so to will our ability to support computer users in doing FM and indeed in any information task that similarly features labels and categories. Given the ubiquity of such contexts in today's world, this is a promising research direction, which we now better understand and are better equipped to study further.

Bibliography

Adrian, B., Klinkigt, M., Maus, H., & Dengel, A. (2009). Using idocument for document categorization in nepomuk social semantic desktop. In *I-semantics* (pp. 638–643).

Adrian, B., Sauermann, L., & Roth-Berghofer, T. (2007). Contag: A semantic tag recommendation system. *Proceedings of I-Semantics*, *7*, 297–304.

Agarawala, A., & Balakrishnan, R. (2006). Keepin'it real: pushing the desktop metaphor with physics, piles and the pen. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (p. 1283–1292).

Agrawal, N., Bolosky, W. J., Douceur, J. R., & Lorch, J. R. (2007). A five-year study of file-system metadata. *ACM Transactions on Storage (TOS)*, *3*(3), 9.

Akin, O., Baykan, C., & Rao, D. R. (1987). Structure of a directory space: A case study with a UNIX operating system. *International journal of man-machine studies*, *26*(3), 361–382.

Albadri, N., Watson, R., & Dekeyser, S. (2016). TreeTags: bringing tags to the hierarchical file system. In *Proceedings of the Australasian Computer Science Week Multiconference* (p. 21).

Altom, T., Buher, M., Downey, M., & Faiola, A. (2004). Using 3D landscapes to navigate file systems: the MountainView interface. In *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on* (p. 645–649).

Baker, M. G., Hartman, J. H., Kupfer, M. D., Shirriff, K. W., & Ousterhout, J. K. (1991). Measurements of a distributed file system. In *ACM SIGOPS Operating Systems Review* (Vol. 25, p. 198–212).

Barreau, D. (1995). Context as a factor in personal information management systems.

Journal of the American Society for Information Science, *46*(5), 327–339.

Barreau, D., & Nardi, B. A. (1995). Finding and reminding: file organization from the desktop. *ACM SigChi Bulletin*, *27*(3), 39–43.

Bass, J. (2013). A pim perspective: leveraging personal information management research in the archiving of personal digital records. *Archivaria*, *75*.

Bauer, D., Fastrez, P., & Hollan, J. (2005). Spatial tools for managing personal information collections. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on* (p. 104b–104b).

Benn, Y., Bergman, O., Glazer, L., Arent, P., Wilkinson, I. D., Varley, R., & Whittaker, S. (2015). Navigating through digital folders uses the same brain structures as real world navigation. *Scientific reports*, *5*.

Bennett, J. M., Bauer, M. A., & Kinchlea, D. (1991). Characteristics of files in NFS environments. In *Proceedings of the 1991 ACM SIGSMALL/PC symposium on Small systems* (p. 33–40).

Bergman, O. (2012). The User-Subjective Approach to Personal Information Management: From Theory to Practice. In M. Zacarias & J. V. Oliveira (Eds.), *Human-Computer Interaction: The Agency Perspective* (Vol. 396, p. 55–81). Springer Berlin Heidelberg.

Bergman, O. (2013). Variables for personal information management research. In *Aslib Proceedings* (Vol. 65, p. 1–1).

Bergman, O., Beyth-Marom, R., & Nachmias, R. (2003). The user-subjective approach to personal information management systems. *Journal of the American Society for Information Science and Technology*, *54*(9), 872–878.

Bergman, O., Beyth-Marom, R., & Nachmias, R. (2006). The project fragmentation problem in personal information management. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (p. 271–274).

Bergman, O., Beyth-Marom, R., & Nachmias, R. (2008). The user-subjective approach to personal information management systems design: Evidence and implementations. *Journal of the American Society for Information Science and Technology*, *59*(2),

235–246.

Bergman, O., Beyth-Marom, R., Nachmias, R., Gradovitch, N., & Whittaker, S. (2008). Improved search engines and navigation preference in personal information management. *ACM Transactions on Information Systems (TOIS)*, *26*(4), 20.

Bergman, O., Elyada, O., Dvir, N., Vaitzman, Y., & Ami, A. B. (2014). Spotting the Latest Version of a File with Old'nGray. *Interacting with Computers*, iwu018.

Bergman, O., Gradovitch, N., Bar-Ilan, J., & Beyth-Marom, R. (2013a). Folder versus tag preference in personal information management. *Journal of the American Society for Information Science and Technology*, *64*(10), 1995–2012.

Bergman, O., Gradovitch, N., Bar-Ilan, J., & Beyth-Marom, R. (2013b, November). Tagging Personal Information: A Contrast between Attitudes and Behavior. In *ASIST 2013.* Montréal, Quebec, Canada.

Bergman, O., Tene-Rubinstein, M., & Shalom, J. (2013). The use of attention resources in navigation versus search. *Personal and Ubiquitous Computing*, *17*(3), 583–590.

Bergman, O., Tucker, S., Beyth-Marom, R., Cutrell, E., & Whittaker, S. (2009). It's not that important: demoting personal information of low subjective importance using GrayArea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 269–278).

Bergman, O., Whittaker, S., & Falk, N. (2014). Shared files: The retrieval perspective. *Journal of the Association for Information Science and Technology*, *65*(10), 1949–1963. Retrieved from `http://dx.doi.org/10.1002/asi.23147` doi: 10.1002/asi.23147

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

Bergman, O., Whittaker, S., Sanderson, M., Nachmias, R., & Ramamoorthy, A. (2012). How do we find personal files?: the effect of os, presentation & depth on file navigation. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems* (p. 2977–2980).

Berlin, L. M., Jeffries, R., O'Day, V. L., Paepcke, A., & Wharton, C. (1993). Where did you put it? Issues in the design and use of a group memory. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (p. 23–30).

Blanc-Brude, T., & Scapin, D. L. (2007). What do people recall about their documents?: implications for desktop search tools. In *Proceedings of the 12th international conference on Intelligent user interfaces* (p. 102–111).

Bloehdorn, S., Görlitz, O., Schenk, S., & Völkel, M. (2006). Tagfs – tag semantics for hierarchical file systems. In *Proceedings of the 6th International Conference on Knowledge Management (I-KNOW 06), Graz, Austria* (Vol. 8).

Boardman, R., & Sasse, M. A. (2004). Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (p. 583–590).

Bondarenko, O., & Janssen, R. (2005). Documents at hand: Learning from paper to improve digital technologies. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (p. 121–130).

Bowman, C. M., Dharap, C., Baruah, M., Camargo, B., & Potti, S. (1994, June). A File System for Information Management. In *Proceedings of the Conference on Intelligent Information Management Systems.* Citeseer.

Brinegar, J., & Capra, R. (2010). Understanding personal digital music collections. *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–2.

Brinegar, J., & Capra, R. (2011). Managing music across multiple devices and computers. In *Proceedings of the 2011 iconference* (pp. 489–495).

Brostoff, S., Sasse, M. A., Chadwick, D., Cunningham, J., Mbanaso, U., & Otenko, S. (2005). 'r-what?'development of a role-based access control policy-writing tool for e-scientists. *Software: Practice and Experience*, *35*(9), 835–856.

Bruce, H., Jones, W., & Dumais, S. (2004). Information behaviour that keeps found things found. *Information Research*, *10*(1), paper 207.

Bryman, A. (2012). *Social research methods.* Oxford university press.

Capra, R. (2009). A survey of personal information management practices. Vancouver, British Columbia, Canada.

Capra, R., & Perez-Quinones, M. (2006). Factors and evaluation of refinding behaviors. In *SIGIR 2006 Workshop on Personal Information Management* (p. 10–11).

Capra, R., Pinney, M., & Perez-Quinones, M. (2005). *Refinding is not finding again.* (Tech. Rep. No. TR-05–10). Blacksburg, Virginia: Computer Science Department, Virginia Tech.

Capra, R., Vardell, E., & Brennan, K. (2014, October 31 - November 5). File Synchronization and Sharing: User Practices and Challenges. In *77th Annual ASIS&T Meeting.* Seattle, WA, USA.

Carroll, J. M. (1982). Creative names for personal files in an interactive computing environment. *International Journal of Man-Machine Studies*, *16*(4), 405–438.

Case, D. O. (1986). Collection and organization of written information by social scientists and humanists: a review and exploratory study. *Journal of Information Science*, *12*(3), 97–104.

Case, D. O. (1991). Conceptual organization and retrieval of text by historians: The role of memory and metaphor. *JASIS*, *42*(9), 657–668.

Chen, C., & Rada, R. (1996). Interacting with hypertext: A meta-analysis of experimental studies. *Human-computer interaction*, *11*(2), 125–156.

Chernov, S., Demartini, G., Herder, E., Kopycki, M., & Nejdl, W. (2008). Evaluating personal information management using an activity logs enriched desktop dataset. In *Proceedings of 3rd personal information management workshop (pim 2008), florence, italy* (Vol. 155).

Chirita, P.-A., Costache, S., Nejdl, W., & Paiu, R. (2006). Beagle++: Semantically enhanced searching and ranking on the desktop. In *European semantic web conference* (pp. 348–362).

Civan, A., Jones, W., Klasnja, P., & Bruce, H. (2008). Better to organize personal information by folders or by tags?: The devil is in the details. In (Vol. 45, p. 1–13).

Wiley Online Library.

Cole, B. (2005). Search engines tackle the desktop. *Computer*, *38*(3), 14–17.

Cole, I. (1982). Human aspects of office filing: Implications for the electronic office. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 26, p. 59–63).

Corbató, F. J., Merwin-Daggett, M., & Daley, R. C. (1962). An experimental time-sharing system. In *Proceedings of the May 1-3, 1962, spring joint computer conference* (p. 335–344).

Cox, L. P., Murray, C. D., & Noble, B. D. (2002). Pastiche: Making backup cheap and easy. *ACM SIGOPS Operating Systems Review*, *36*(SI), 285–298.

Cushing, A. L. (2013). "It's stuff that speaks to me": Exploring the characteristics of digital possessions. *Journal of the American Society for Information Science and Technology*, *64*(8), 1723–1734.

Cutrell, E. (2006). Search User Interfaces for PIM. In *Personal Information Management, SIGIR 2006 Workshop* (p. 32–35).

Cutrell, E., Dumais, S. T., & Teevan, J. (2006). Searching to eliminate personal information management. *Communications of the ACM*, *49*(1), 58–64.

Cutrell, E., Robbins, D., Dumais, S., & Sarin, R. (2006). Fast, flexible filtering with phlat. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (p. 261–270).

De Chiara, R., Erra, U., & Scarano, V. (2003). VENNFS: A Venn-diagram file manager. In *2013 17th International Conference on Information Visualisation* (p. 120–120).

Dearman, D., & Pierce, J. S. (2008). It's on my other computer!: computing with multiple devices. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (p. 767–776).

De Chiara, R., Erra, U., & Scarano, V. (2003). Vennfs: A venn-diagram file manager. In *Iv* (pp. 120–126).

Dinneen, J. D., Odoni, F., Frissen, I., & Julien, C.-A. (2016). Cardinal: novel software for studying file management behaviour. In *Asis&t 2016: Proceedings of the 79th*

annual meeting of the association for information science & technology (Vol. 53).

Dinneen, J. D., Odoni, F., & Julien, C.-A. (2016). Towards a desirable data collection tool for studying long-term pim. In *Personal information management workshop at chi '16: Acm conference on human factors in computing systems.*

Dittrich, J.-P., & Salles, M. A. V. (2006). iDM: A unified and versatile data model for personal dataspace management. In *Proceedings of the 32nd international conference on Very large data bases* (p. 367–378).

Dong, X. L., & Halevy, A. (2005). A platform for personal information management and integration. In *Proceedings of VLDB 2005 PhD Workshop* (p. 26).

Douceur, J. R., & Bolosky, W. J. (1999). A large-scale study of file-system contents. *ACM SIGMETRICS Performance Evaluation Review*, *27*(1), 59–70.

Dourish, P., Edwards, W. K., LaMarca, A., Lamping, J., Petersen, K., Salisbury, M., ... Thornton, J. (2000). Extending document management systems with user-specific active properties. *ACM Transactions on Information Systems (TOIS)*, *18*(2), 140–170.

Dourish, P., Edwards, W. K., LaMarca, A., & Salisbury, M. (1999a). Presto: an experimental architecture for fluid interactive document spaces. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *6*(2), 133–161.

Dourish, P., Edwards, W. K., LaMarca, A., & Salisbury, M. (1999b). Using properties for uniform interaction in the Presto document system. In *Proceedings of the 12th annual ACM symposium on User interface software and technology* (p. 55–64).

Dourish, P., Lamping, J., & Rodden, T. (1999). Building bridges: customisation and mutual intelligibility in shared category management. In *Proceedings of the international ACM SIGGROUP conference on Supporting group work* (p. 11–20).

Downey, A. B. (2001). The structural cause of file size distributions. In *Modeling, analysis and simulation of computer and telecommunication systems, 2001. proceedings. ninth international symposium on* (pp. 361–370).

Dragunov, A. N., Dietterich, T. G., Johnsrude, K., McLaughlin, M., Li, L., & Herlocker, J. L. (2005). TaskTracer: a desktop environment to support multi-tasking knowl-

edge workers. In *Proceedings of the 10th international conference on Intelligent user interfaces* (p. 75–82).

Ducheneaut, N., & Bellotti, V. (2001). E-mail as habitat: an exploration of embedded personal information management. *interactions*, *8*(5), 30–38.

Dumais, S., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R., & Robbins, D. C. (2003). Stuff I've seen: a system for personal information retrieval and re-use. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (p. 72–79).

Evans, K. M., & Kuenning, G. H. (2002). A study of irregularities in file-size distributions. In *Proceedings of the 2002 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS).*

Fastrez, P., & Jacques, J. (2015). Managing references by filing and tagging: an exploratory study of personal information management by social scientists. In *17th international conference, hci international 2015.*

Ferguson, R. D. (2016). Ethics and the long-term management of personal information. In *Position papers at 2016 pim workshop at chi '16, paper 7.*

Fertig, S., Freeman, E., & Gelernter, D. (1996a). Lifestreams: an alternative to the desktop metaphor. In *Conference companion on Human factors in computing systems* (p. 410–411).

Fertig, S., Freeman, E., & Gelernter, D. (1996b). "finding and reminding" reconsidered. *ACM SIGCHI Bulletin*, *28*(1), 66–69.

File, T., & Ryan, C. (2014). Computer and internet use in the united states: 2013. *American Community Survey Reports*.

Fisher, K. E., Erdelez, S., & McKechnie, L. (2005). *Theories of information behavior.* Information Today, Inc.

Fitchett, S., & Cockburn, A. (2015). An empirical characterisation of file retrieval. *International Journal of Human-Computer Studies*.

Fitchett, S., Cockburn, A., & Gutwin, C. (2013). Improving navigation-based file retrieval. In *Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems (p. 2329–2338).

Fitchett, S., Cockburn, A., & Gutwin, C. (2014). Finder highlights: field evaluation and design of an augmented file browser. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (p. 3685–3694).

Floridi, L. (2010). *Information: A very short introduction*. Oxford University Press.

Ford, N., Wilson, T., Foster, A., Ellis, D., & Spink, A. (2002). Information seeking and mediated searching. Part 4. Cognitive styles in information seeking. *Journal of the American Society for Information Science and Technology*, *53*(9), 728–735.

Fourie, I. (2011). Librarians alert: how can we exploit what is happening with personal information management (PIM), reference management and related issues? *Library Hi Tech*, *29*(3), 550–556.

Freeman, E., & Gelernter, D. (1996). Lifestreams: A storage model for personal data. *ACM SIGMOD Record*, *25*(1), 80–86.

Frohlich, D. M. (1997). Direct manipulation and other lessons. In *Handbook of human-computer interaction* (pp. 463–488). Elsevier Science BV.

Gao, Q. (2011). An empirical study of tagging for personal information organization: Performance, workload, memory, and consistency. *International Journal of Human-Computer Interaction*, *27*(9), 821–863.

Gemmell, J., Bell, G., Lueder, R., Drucker, S., & Wong, C. (2002). Mylifebits: fulfilling the memex vision. In *Proceedings of the tenth acm international conference on multimedia* (pp. 235–238).

Ghorashi, S., & Jensen, C. (2012). Leyline: Provenance-based search using a graphical sketchpad. In *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval* (p. 2).

Gifford, D. K., Jouvelot, P., & Sheldon, M. A. (1991). Semantic file systems. In *ACM SIGOPS Operating Systems Review* (Vol. 25, p. 16–25).

Gill, S. L. (1988). *File management and information retrieval systems: a manual for managers and technicians*. Libraries Unlimited, Inc.

Gonçalves, D. J., & Jorge, J. A. (2003a). Analyzing personal document spaces. In

Proceedings of Human-Computer Interaction International (Vol. 24).

Gonçalves, D. J., & Jorge, J. A. (2003b). An empirical study of personal document spaces. In *Interactive Systems. Design, Specification, and Verification* (p. 46–60). Springer.

Gonçalves, D. J., & Jorge, J. A. (2006). Quill: a narrative-based interface for personal document retrieval. In *Chi'06 extended abstracts on human factors in computing systems* (pp. 327–332).

Gonçalves, D. J., & Jorge, J. A. (2008a). In search of personal information: narrative-based interfaces. In *Proceedings of the 13th international conference on intelligent user interfaces* (pp. 179–188).

Gonçalves, D. J., & Jorge, J. A. (2008b). Now, It's Personal! Evaluating PIM Retrieval Tools. In *PIM workshop, CHI 2008* (p. 5–6).

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: personality judgments based on offices and bedrooms. *Journal of personality and social psychology*, *82*(3), 379.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, *37*(6), 504–528.

Gwizdka, J., & Chignell, M. (2007). Individual Differences. In (p. 206–220). Seattle, WA: University of Washington Press.

Gyllstrom, K. (2009). *Enriching personal information management with document interaction histories* (Unpublished doctoral dissertation). University of North Carolina at Chapel Hill.

Halasz, F., & Moran, T. P. (1982). Analogy considered harmful. In *Proceedings of the 1982 conference on Human factors in computing systems* (p. 383–386).

Haller, H., & Abecker, A. (2010). imapping: a zooming user interface approach for personal and semantic knowledge management. In *Proceedings of the 21st acm conference on hypertext and hypermedia* (pp. 119–128).

Handschuh, S., Möller, K., & Groza, T. (2007). The nepomuk project-on the way to the social semantic desktop. In *I-semantics 2007.* Graz, Austria.

Hardof-Jaffe, S., Hershkovitz, A., Abu-Kishk, H., Bergman, O., & Nachmias, R. (2009a). How do students organize personal information spaces? *International Working Group on Educational Data Mining*.

Hardof-Jaffe, S., Hershkovitz, A., Abu-Kishk, H., Bergman, O., & Nachmias, R. (2009b). Students' organization strategies of personal information space. *Journal of Digital Information*, *10*(5).

Hardy, D. R., & Schwartz, M. F. (1993). Essence: A Resource Discovery System Based on Semantic File Indexing. In *USENIX Winter* (p. 361–374).

Hariri, N., Asadi, M., & Mansourian, Y. (2014). The impact of users' verbal/imagery cognitive styles on their Web search behavior. *Aslib Journal of Information Management*, *66*(4), 401–423.

Harper, R., Lindley, S., Thereska, E., Banks, R., Gosset, P., Smyth, G., ... Whitworth, E. (2013). What is a File? In *Proceedings of the 2013 conference on Computer supported cooperative work* (p. 1125–1136).

Harter, T., Dragga, C., Vaughn, M., Arpaci-Dusseau, A. C., & Arpaci-Dusseau, R. H. (2012). A file is not a file: understanding the i/o behavior of apple desktop applications. *ACM Transactions on Computer Systems (TOCS)*, *30*(3), 10.

Hawkins, D. (Ed.). (2013). *Personal archiving: preserving our digital heritage*. Information Today, Inc.

Hecht, M. (1985). *File and data-base management programs for the IBM PC*. John Wiley & Sons, Inc.

Henderson, S. (2005). Genre, task, topic and time: facets of personal digital document management. In *Proceedings of the 6th ACM SIGCHI New Zealand chapter's international conference on Computer-human interaction: making CHI natural* (p. 75–82).

Henderson, S. (2011). Document duplication: How users (struggle to) manage file copies and versions. *Proceedings of the American Society for Information Science and Technology*, *48*(1), 1–10.

Henderson, S., & Srinivasan, A. (2009). An empirical analysis of personal digital docu-

ment structures. In *Human Interface and the Management of Information. Designing Information Environments* (p. 394–403). Springer.

Henderson, S., & Srinivasan, A. (2011). Filing, piling & structuring: strategies for personal document management. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on* (p. 1–10).

Hicks, B. J., Dong, A., Palmer, R., & McAlpine, H. C. (2008). Organizing and managing personal electronic files: A mechanical engineer's perspective. *ACM Transactions on Information Systems (TOIS)*, *26*(4), 23.

Hirakawa, M., Mizumoto, S., Yoshitaka, A., & Ichikawa, T. (1998). A situation-sensitive interface for the management of personal documents. In *Multimedia Software Engineering, 1998. Proceedings. International Workshop on* (p. 96–103).

Hjørland, B., & Hjørland, B. (2005). Empiricism, rationalism and positivism in library and information science. *Journal of Documentation*, *61*(1), 130–155.

Hjørland, B. (2002). Domain analysis in information science: eleven approaches-traditional as well as innovative. *Journal of documentation*, *58*(4), 422–462.

Hsieh, J., Chen, C., Lin, I., & Sun, C. (2008). A web-based tagging tool for organizing personal documents on pcs. In *International Conference of Computer-Human Interaction.*

Hui, Y. (2012). What is a digital object? *Metaphilosophy*, *43*(4), 380–395.

Huvila, I., Eriksen, J., Häusner, E.-M., & Jansson, I.-M. (2014). Continuum thinking and the contexts of personal information management. *Information Research: An International Electronic Journal*, *19*(1), n1.

Jensen, C., Lonsdale, H., Wynn, E., Cao, J., Slater, M., & Dietterich, T. G. (2010). The life and times of files and information: a study of desktop provenance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 767–776).

Jeuris, S., Houben, S., & Bardram, J. (2014). Laevo: a temporal desktop interface for integrated knowledge work. In *Proceedings of the 27th annual ACM symposium on User interface software and technology* (p. 679–688).

Johnson, B., & Shneiderman, B. (1991). Tree-maps: A space-filling approach to the visualization of hierarchical information structures. In *Visualization, 1991. visualization'91, proceedings., ieee conference on* (pp. 284–291).

Johnson, J. (1987). How faithfully should the electronic office simulate real one? *ACM SIGCHI Bulletin*, *19*(2), 21–25.

Johnson, M. L., Bellovin, S. M., Reeder, R. W., & Schechter, S. E. (2009). Laissez-faire file sharing. In *New Security Paradigms Workshop* (Vol. 2009).

Jones, W. (2007a). *Keeping Found Things Found: The Study and Practice of Personal Information Management*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Jones, W. (2007b). Personal information management. *Annual review of information science and technology*, *41*(1), 453–504.

Jones, W., Bellotti, V., Capra, R., Dinneen, J. D., Mark, G., Marshall, C. C., ... Van Kleek, M. (2016). For richer, for poorer, in sickness or in health...: The long-term management of personal information. In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems* (pp. 3508–3515).

Jones, W., Bruce, H., & Dumais, S. (2001). Keeping found things found on the web. In *Proceedings of the tenth international conference on Information and knowledge management* (p. 119–126).

Jones, W., Dinneen, J. D., Capra, R., Pérez-Quiñones, M., & Diekema, A. R. (2015). Personal Information Management (PIM). In J. McDonald & M. Levine-Clark (Eds.), *Encyclopedia of Library and Information Sciences* (4th ed.). CRC Press.

Jones, W., Dinneen, J. D., Capra, R., Pérez-Quiñones, M., & Diekema, A. R. (2017). Personal Information Management (PIM). In J. McDonald & M. Levine-Clark (Eds.), *Encyclopedia of Library and Information Sciences* (4th ed.). CRC Press.

Jones, W., Dumais, S., & Bruce, H. (2002). Once found, what then? A study of "keeping" behaviors in the personal use of Web information. In (Vol. 39, p. 391–402). Wiley Online Library.

Jones, W., Hou, D., Sethanandha, B. D., Bi, S., & Gemmell, J. (2010). Planz to put our digital information in its place. In *Chi'10 extended abstracts on human factors in*

computing systems (pp. 2803–2812).

Jones, W., Phuwanartnurak, A. J., Gill, R., & Bruce, H. (2005). Don't take my folders away!: organizing personal information to get ghings done. In *CHI'05 extended abstracts on Human factors in computing systems* (p. 1505–1508).

Jones, W., Thorsteinson, C., Thepvongsa, B., & Garrett, T. (2016). Making it real: Towards practical progress in the management of personal information. In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems* (pp. 571–582).

Jones, W., Wenning, A., & Bruce, H. (2014). How do people re-find files, emails and web pages? *iConference 2014 Proceedings*.

Julien, C.-A., Asadi, B., Dinneen, J. D., & Shu, F. (2016). Library of congress subject heading (lcsh) browsing and natural language searching. In *Asis&t 2016: Proceedings of the 79th annual meeting of the association for information science & technology* (Vol. 53).

Julien, C.-A., Tirilly, P., Dinneen, J. D., & Guastavino, C. (2013). Reducing Subject Tree Browsing Complexity. *Journal of the American Society for Information Science and Technology*, *64*(11), 2201–2223.

Kalman, Y. M., & Ravid, G. (2015). Filing, piling, and everything in between: The dynamics of E-mail inbox management. *Journal of the Association for Information Science and Technology*, *66*(12), 2540–2552.

Kaptelinin, V. (1996). Creating computer-based work environments: an empirical study of Macintosh users. In *Proceedings of the 1996 ACM SIGCPR/SIGMIS conference on Computer personnel research* (p. 360–366).

Kaptelinin, V. (2003). Umea: translating interaction histories into project contexts. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 353–360).

Kaye, J., Vertesi, J., Avery, S., Dafoe, A., David, S., Onaga, L., ... Pinch, T. (2006). To have and to hold: exploring the personal archive. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (p. 275–284).

Kelly, D. (2006). Evaluating personal information management behaviors and tools. *Communications of the ACM*, *49*(1), 84–86.

Khoo, C., Luyt, B., Ee, C., Osman, J., Lim, H.-H., & Yong, S. (2007). How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour. *Information Research*, *11*(2), 12–2.

Kieras, D. E. (1999). A guide to goms model usability evaluation using gomsl and glean3. *University of Michigan*(313).

Kim, J., & Croft, W. B. (2010). Ranking using multiple document types in desktop search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (p. 50–57).

Kim, J., Croft, W. B., Smith, D., & Bakalov, A. (2011). Evaluating an associative browsing model for personal information. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (p. 647–652).

Kinley, K., Tjondronegoro, D., Partridge, H., & Edwards, S. (2014). Modeling users' web search behavior and their cognitive styles. *Journal of the Association for Information Science and Technology*, *65*(6), 1107–1123.

Kljun, M., Mariani, J., & Dix, A. (2015a). Toward understanding short-term personal information preservation: A study of backup strategies of end users. *Journal of the Association for Information Science and Technology*.

Kljun, M., Mariani, J., & Dix, A. (2015b). Transference of PIM research prototype concepts to the mainstream: successes or failures. *Interacting with Computers*, 73–98. doi: 10.1093/iwc/iwt059

Kobsa, A. (2004). User experiments with tree visualization systems. In *Information visualization, 2004. infovis 2004. ieee symposium on* (pp. 9–16).

Kozhevnikov, M. (2007). Cognitive styles in the context of modern psychology: toward an integrated framework of cognitive style. *Psychological bulletin*, *133*(3), 464.

Krishnan, A., & Jones, S. (2005). TimeSpace: activity-based temporal visualisation of personal information spaces. *Personal and Ubiquitous Computing*, *9*(1), 46–65.

Kwasnik, B. H. (1991). The importance of factors that are not document attributes in the

organisation of personal documents. *Journal of documentation*, *47*(4), 389–398.

Lansdale, M. W. (1988). The psychology of personal information management. *Applied ergonomics*, *19*(1), 55–66.

Lee, B. L., & Bederson, B. B. (2003, November). Favorite folders: A configurable, scalable file browser. In *ACM Symposium on User Interface Software and Technology (UIST) 2003.* Vancouver, BC, Canada.

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, *51*(5), 341–352.

Liu, G., & Feng, L. (2016). A Method to Support Difficult Re-finding Tasks. *arXiv preprint arXiv:1601.07273*.

Lyman, P., Varian, H. R., Swearingen, K., Charles, P., Good, N., Jordan, L., & Pal, J. (2003). *How much information?* (Tech. Rep.). Berkeley, CA: University of California, Berkeley. (Online; accessed 19-July-2008)

Ma, S., & Wiedenbeck, S. (2009). File management with hierarchical folders and tags. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems* (p. 3745–3750).

Mackenzie, M. L. (2000). The personal organization of electronic mail messages in a business environment: An exploratory study. *Library & information science research*, *22*(4), 405–426.

Malone, T. W. (1983). How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Information Systems (TOIS)*, *1*(1), 99–112.

Manber, U. (1994). Finding similar files in a large file system. In *Usenix winter* (Vol. 94, pp. 1–10).

Mander, R., Salomon, G., & Wong, Y. Y. (1992). A "pile" metaphor for supporting casual organization of information. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (p. 627–634).

Mark, G., & Prinz, W. (1997). What happened to our document in the shared workspace? The need for Groupware conventions. In *Human-Computer Interaction INTER-*

ACT'97 (p. 413–420).

Marsden, G., & Cairns, D. E. (2003). Improving the usability of the hierarchical file system. In *Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology* (p. 122–129).

Marshall, C. C., Bly, S., & Brun-Cottan, F. (2006). The long term fate of our digital belongings: Toward a service model for personal archives. In *Archiving Conference* (Vol. 2006, p. 25–30).

Marshall, C. C., McCown, F., & Nelson, M. L. (2007). Evaluating personal archiving strategies for Internet-based information. In *Archiving Conference* (Vol. 2007, p. 151–156).

Marshall, C. C., & Tang, J. C. (2012). That syncing feeling: early user experiences with the cloud. In *Proceedings of the designing interactive systems conference* (pp. 544–553).

Marshall, C. C., Wobber, T., Ramasubramanian, V., & Terry, D. B. (2012). Supporting research collaboration through bi-level file synchronization. In *Proceedings of the 17th acm international conference on supporting group work* (pp. 165–174).

Massey, C., TenBrook, S., Tatum, C., & Whittaker, S. (2014). Pim and personality: what do our personal file systems say about us? In *Proceedings of the 32nd annual acm conference on human factors in computing systems* (p. 3695–3704).

Meho, L. I., & Tibbo, H. R. (2003). Modeling the information-seeking behavior of social scientists: Ellis's study revisited. *Journal of the American society for Information Science and Technology*, *54*(6), 570–587.

Merčun, T., & Žumer, M. (2013). User perception of 4 hierarchical layouts. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–3.

Millen, D. R., Yang, M., Whittaker, S., & Feinberg, J. (2007). Social bookmarking and exploratory search. In *Ecscw 2007* (pp. 21–40). Springer.

Mosweunyane, G., Carr, L., & Gibbins, N. (2011). A Tag-Like, Linked Navigation Approach for Retrieval and Discovery of Desktop Documents. In *Digital Information*

and Communication Technology and Its Applications (p. 692–706). Springer.

Nardi, B. A., Anderson, K., & Erickson, T. (1995). Filing and finding computer files..

Nelson, T. H. (1965). Complex information processing: a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th national conference* (p. 84–100).

Nelson, T. H. (2000). Xanalogical structure. *Needed Now More Than Ever: Parallel Documents, Deep Links to Content, Deep Visioning, and Deep Re-Use." Xanalogical Structure*.

Odom, W., Zimmerman, J., & Forlizzi, J. (2011). Teenagers and their virtual possessions: design opportunities and issues. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1491–1500).

Oh, K. E. (2017). Types of personal information categorization: Rigid, fuzzy, and flexible. *Journal of the Association for Information Science and Technology*.

Oleksik, G., Wilson, M. L., Tashman, C., Mendes Rodrigues, E., Kazai, G., Smyth, G., ... Jones, R. (2009). Lightweight tagging expands information and activity management practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 279–288).

Otopah, F. O., & Dadzie, P. (2013). Personal information management practices of students and its implications for library services. In *Aslib proceedings* (Vol. 65, pp. 143–160).

Ousterhout, J. K., Da Costa, H., Harrison, D., Kunze, J. A., Kupfer, M., & Thompson, J. G. (1985). *A trace-driven analysis of the unix 4.2 bsd file system* (Vol. 19) (No. 5). ACM.

O'hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122.

Paré, F.-X. (2007). The many facets of document importance: A case study of office workers. In *Canadian Association of Information Science Annual Conference, Montreal, Quebec.* (Vol. 19).

Paré, F.-X. (2011). *Personal information management among office support staff in a*

university environment: an exploratory study (Unpublished doctoral dissertation). McGill University, Montreal.

Parkin, T., & Robinson, J. (1992). Analysis of lognormal data. In *Advances in soil science* (pp. 193–235). Springer.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information.* Oxford University Press.

Plaisant, C., Grosjean, J., & Bederson, B. B. (2002). Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation. In *Information visualization, 2002. infovis 2002. ieee symposium on* (pp. 57–64).

Prinz, W., & Zaman, B. (2005). Proactive support for the organization of shared workspaces using activity patterns and content analysis. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work* (p. 246–255).

Quan, D., Bakshi, K., Huynh, D., & Karger, D. R. (2003). User interfaces for supporting multiple categorization. In *Proceedings of INTERACT* (p. 228–235).

Ramer, S. L. (2005). Site-ation pearl growing: methods and librarianship history and theory. *Journal of the Medical Library Association*, *93*(3), 397.

Ravasio, P., Schär, S. G., & Krueger, H. (2004). In pursuit of desktop evolution: User problems and practices with modern desktop systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *11*(2), 156–180.

Ravasio, P., & Tscherter, V. (2007). Users' theories of the desktop metaphor, or why we should seek metaphor-free interfaces. *Beyond the desktop metaphor: designing integrated digital work environments. MIT Press, Cambridge, MA*, 265–294.

Riding, R., & Cheema, I. (1991). Cognitive styles—an overview and integration. *Educational psychology*, *11*(3-4), 193–215.

Riding, R. J. (1997). On the nature of cognitive style. *Educational Psychology*, *17*(1-2), 29–49.

Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., & Van Dantzich, M. (1998). Data mountain: using spatial memory for document management. In *Proceedings of the 11th annual ACM symposium on User interface software and*

technology (p. 153–162).

Robertson, G. G., Mackinlay, J. D., & Card, S. K. (1991). Cone trees: animated 3d visualizations of hierarchical information. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 189–194).

Rodden, K., & Wood, K. R. (2003). How do people manage their digital photographs? In *Proceedings of the SIGCHI conference on Human factors in computing systems* (p. 409–416).

Rode, J., Johansson, C., DiGioia, P., Nies, K., Nguyen, D. H., Ren, J., ... Redmiles, D. (2006). Seeing further: extending visualization as a basis for usable security. In *Proceedings of the second symposium on usable privacy and security* (pp. 145–155).

Roselli, D. S., Lorch, J. R., & Anderson, T. E. (2000). A Comparison of File System Workloads. In *USENIX Annual Technical Conference, General Track* (p. 41–54).

Rowley, J. E., & Hartley, R. J. (2008). *Organizing knowledge: an introduction to managing access to information.* Ashgate Publishing, Ltd.

Sajedi, A., Afzali, S. H., & Zabardast, Z. (2012). Can you retrieve a file on the computer in your first attempt? think to a new file manager for multiple categorization of your personal information. In *6th international workshop on personal information management.*

Salmon, B. W. (2009). *Putting home data management into perspective* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

Satyanarayanan, M. (1981). A study of file sizes and functional lifetimes. In *ACM SIGOPS Operating Systems Review* (Vol. 15, p. 96–108).

Sauermann, L., Bernardi, A., & Dengel, A. (2005). Overview and outlook on the semantic desktop. In *Proceedings of the 2005 international conference on semantic desktop workshop: Next generation information management d collaboration infrastructure-volume 175* (pp. 74–91).

Sauermann, L., Grimnes, G. A., Kiesel, M., Fluit, C., Maus, H., Heim, D., ... Dengel, A. (2006). Semantic desktop 2.0: The gnowsis experience. In *International semantic web conference* (pp. 887–900).

Schaffer, D., & Greenberg, S. (1993). Sifting through hierarchical information. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems* (p. 173–174).

Seebach, P. (2001). *The cranky user: The Principle of Least Astonishment.* Agosto de.

Seltzer, M. I., & Murphy, N. (2009). Hierarchical File Systems Are Dead. In *HotOS.*

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets* (Unpublished doctoral dissertation). University of Pittsburgh.

Shneiderman, B., & Plaisant, C. (1994). The future of graphic user interfaces: Personal role managers. In *Bcs hci* (pp. 3–8).

Siddiqui, S., & Turley, D. (2006). Extending the self in a virtual world. *NA-Advances in Consumer Research Volume 33*.

Sienknecht, T. F., Friedrich, R. J., Martinka, J. J., & Friedenbach, P. M. (1994). The implications of distributed data in a commercial environment on the design of hierarchical storage management. *Performance Evaluation*, *20*(1), 3–25.

Sinha, D., & Basu, A. (2012a). Design and evaluation of a cognition aware file browser for users in rural india. In *Perception and machine intelligence* (pp. 129–136). Springer.

Sinha, D., & Basu, A. (2012b). Gardener: A file browser assistant to help users maintaining semantic folder hierarchy. In *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on* (p. 1–6).

Sinha, D., & Basu, A. (2012c). Natural arrangement: A novel and intuitive perspective on filesystem re-organization. In *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on* (p. 1–4).

Smiraglia, R. P. (2002). The progress of theory in knowledge organization. *Library trends*, *50*(3), 330–349.

Smith, A. J. (1981). Analysis of long term file reference patterns for application to file migration algorithms. *Software Engineering, IEEE Transactions on*(4), 403–417.

Smith, G., Czerwinski, M., Meyers, B., Robbins, D., Robertson, G., & Tan, D. S. (2006). Facetmap: A scalable search and browse visualization. *IEEE Transactions on vi-*

sualization and computer graphics, *12*(5), 797–804.

Stasko, J., Catrambone, R., Guzdial, M., & McDonald, K. (2000). An evaluation of space-filling information visualizations for depicting hierarchical structures. *International journal of human-computer studies*, *53*(5), 663–694.

Sternberg, R. (2008). *Cognitive psychology*. Cengage Learning.

Tanenbaum, A. S., Herder, J. N., & Bos, H. (2006). File size distribution on UNIX systems: then and now. *ACM SIGOPS Operating Systems Review*, *40*(1), 100–104.

Tang, J. C., Brubaker, J. R., & Marshall, C. C. (2013). What do you see in the cloud? understanding the cloud-based user experience through practices. In *Ifip conference on human-computer interaction* (pp. 678–695).

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (p. 415–422).

Thai, V., Handschuh, S., & Decker, S. (2008). *IVEA: An information visualization tool for personalized exploratory document collection analysis*. Springer.

Treglown, M. (2000). Embodiment and interface metaphors: Comparing computer filing systems. In *People and computers xiv—usability or else!* (pp. 341–355). Springer.

Trullemans, S., & Signer, B. (2014a). From user needs to opportunities in personal information management: A case study on organisational strategies in cross-media information spaces. In *Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on* (p. 87–96).

Trullemans, S., & Signer, B. (2014b). Towards a Conceptual Framework and Metamodel for Context-Aware Personal Cross-Media Information Management Systems. In *Conceptual Modeling* (p. 313–320). Springer.

Tsianos, N., Germanakos, P., Lekkas, Z., Mourlas, C., & Samaras, G. (2009). Eye-tracking users' behavior in relation to cognitive style within an e-learning environment. In *Advanced Learning Technologies, 2009. ICALT 2009. Ninth IEEE International Conference on* (p. 329–333).

Tungare, M., & Perez-Quinones, M. (2008). Thinking outside the (beige) box: Personal information management beyond the desktop. In *Proceedings of the workshop on personal information management at sigchi 2008* (p. 8).

Turo, D., & Johnson, B. (1992). Improving the visualization of hierarchies with treemaps: design issues and experimentation. In *Proceedings of the 3rd conference on visualization'92* (pp. 124–131).

Vicente, K. J., Hayes, B. C., & Williges, R. C. (1987). Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *29*(3), 349–359.

Vicente, K. J., & Williges, R. C. (1988). Accommodating individual differences in searching a hierarchical file system. *International Journal of Man-Machine Studies*, *29*(6), 647–668.

Vogels, W. (1999). File system usage in Windows NT 4.0. *ACM SIGOPS Operating Systems Review*, *33*(5), 93–109.

Voida, A., Olson, J. S., & Olson, G. M. (2013). Turbulence in the clouds: challenges of cloud-based information work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 2273–2282).

Voida, S., Edwards, W. K., Newman, M. W., Grinter, R. E., & Ducheneaut, N. (2006). Share and share alike: exploring the user interface affordances of file sharing. In *Proceedings of the SIGCHI conference on Human Factors in computing systems* (p. 221–230).

Voida, S., & Greenberg, S. (2009). WikiFolders: augmenting the display of folders to better convey the meaning of files. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 1679–1682).

Voida, S., & Mynatt, E. D. (2009). It feels better than filing: everyday work experiences in an activity-based computing system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 259–268).

Voida, S., Mynatt, E. D., & Edwards, W. K. (2008). Re-framing the desktop interface around the activities of knowledge work. In *Proceedings of the 21st annual acm*

symposium on user interface software and technology (pp. 211–220).

Voit, K., Andrews, K., & Slany, W. (2012a). Creating a comparative environment for pim evaluation. In *PIM12 CSCW 2012 Workshop, Seattle, WA, USA.*

Voit, K., Andrews, K., & Slany, W. (2012b). Tagging might not be slower than filing in folders. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems* (p. 2063–2068).

Whalen, T., Toms, E. G., & Blustein, J. (2008). Information displays for managing shared files. In *Proceedings of the 2nd ACM Symposium on Computer Human Interaction for Management of Information Technology* (p. 5).

Whitham, R., & Cruickshank, L. (2017). The function and future of the folder. *Interacting with Computers*, 1–19.

Whittaker, S. (2011). Personal information management: from information consumption to curation. *Annual review of information science and technology*, *45*(1), 1–62.

Wideroos, K., & Pekkola, S. (2007). Studying utility of personal usage-history: a software tool for enabling empirical research. In *Human-Computer Interaction. Interaction Design and Usability* (p. 976–984). Springer.

Wilcox, R. (2011). *Modern statistics for the social and behavioral sciences: A practical introduction*. CRC press.

Wulf, V. (1997). Storing and retrieving documents in a shared workspace: experiences from the political administration. In *Human-Computer Interaction INTERACT'97* (p. 469–476).

Xie, X., Sonnenwald, D. H., & Fulton, C. (2015). The role of memory in document re-finding. *Library Hi Tech*, *33*(1), 83-102.

Xu, W., Esteva, M., & Jain, S. D. (2010). Visualizing personal digital collections. In *Proceedings of the 10th annual joint conference on digital libraries* (pp. 169–172).

Zhang, H., & Hu, X. (2014). A quantitative comparison on file folder structures of two groups of information workers. In *Digital libraries (jcdl), 2014 ieee/acm joint conference on* (p. 485-486).

Zhang, H., & Twidale, M. (2012). Mine, yours and ours: Using shared folders in personal

information management. *Personal Information Management (PIM)*.

Appendices

Appendix A: sample of raw data

The following is a portion of raw data collected by Cardinal, the software described in the second chapter and utilised in the third. The data here are from a pilot participant but have been reduced for brevity, showing only one hard drive, one folder (node), and one file. Demographic data, a list of installed file managers, questionnaire responses, and time stamps are not included in this example. Comments (#) have been inserted at the ends of some lines for further explanation.

```
"computer_description": {
        "form": "Laptop",
        "use": "Personal AND Work/School",
        "operating_system": "darwin",   # Darwin is the Mac OS X platform name
        "version": "10.10.5"
    },
    "drives": [{
            "disk_code": "/dev/disk1",
            "size": 122880.0, # Figures are in megabytes; this is a '128 GB' drive
            "used": 63488.0,   # This drive is filled roughly half way to capacity
            "free": 59392.0
      }],
    "node_lists": [
        {    # Begin describing folders on the first hard drive encountered
            "1": {    # Begin describing the first folder encountered
                "node_id": "1",    # Each node is given an ID to identify it since names
                     are not stored
                "depth": 0,    # This folder is the root folder at the top of the tree
                "hard_link_duplicate": false,    # This folder is not present in the
                    tree twice via a hard-link
```

```
                "c_time": "2015-11-30 11:06:23",    # This folder was created in
                    November of 2015
                "m_time": "2015-11-30 11:06:23",    # no files or folders have been
                    added or removed since
                "default": true,    # Name matches a list of default folders for Mac OS
                "name_duplicate": false,    # No other folders have the same name
                "name_length": 11,    # The folder name is 11 characters long
                "letters": 9,    # The folder name contains 9 letters
                "numbers": 2,    # and 2 numbers
                "special_chars": 0,
                "white_spaces": 0,
                "hidden_children": 0,    # No hidden folders within this folder
                "unknown_children": 0,    # No inaccessible (e.g. system) folders within
                     this folder
                "children": ["2",],    # IDs of sub-folders in this folder
                "hidden_files": 2,    # There are two hidden files within this folder
                "symlinks": 0,    # There are no symlinks or shortcuts in this folder
                "file_list": [     # A list of files present in this folder.
                    {
                        "file_id": 1,
                        "extension": "pptx",     # This is a Powerpoint file
                        "file_size": 70636,    # File size is in bytes; this file is ~70
                             KB
                        "hard_link_duplicate": false,
                        "name_duplicate": false,
                        "full_name_length": 46, # This includes the extension and
                            separating doct (e.g., ".pptx")
                        "letters": 35,
                        "numbers": 0,
                        "special_chars": 2,
                        "white_spaces": 4,    # This file has four spaces in its name
                        "c_time": "2015-09-19 19:18:01",    # This file was created in
                            September 2015
                        "m_time": "2015-09-19 19:18:01",   # and hasn't been modified
                            since creation
                        "a_time": "2015-12-13 14:26:53"    # but was last accessed in
                            December, 2015
                    }    # additional files would be listed here
            ]    # end file_list
                }    # end description of the first folder, additional folders would be listed
                    next
            } # end the first node_list (hard drive), additional hard drives would be listed
                next
        ]    # end node_lists
```

Appendix B: data collection code

The following is the *walk* module present in Cardinal (i.e., the data collection software described in Chapter 2) at the time data were collected for this thesis. While Cardinal's source code[5] consists of roughly 1,500 lines of code for the back-end and 2,500 for the user interface, the majority of its novel data collection functionality is provided by the roughly 250 lines of code described below. In short, these utilise Python's *scandir* function to walk through the user-specified locations, where they manage files, and record properties of the relevant files and folders[6] encountered along the way. This produces the raw data seen in Appendix A, which is transferred to the researcher for analysis.

```
#!/usr/bin/env python
""" walks specified disk locations, collects metadata about files and folders. """

from config import DEFAULT_FOLDERS
from classes import File, Node

import sys
import os
import datetime
print('python version ', sys.version)
try:
    from scandir import walk as scandir
    print('found scandir.walk, will use that')
except:
    from os import walk as scandir
    print('didnt find scandir.walk, will use os.walk')

if sys.platform in ['Windows',  'win32']:
    import ctypes



__author__ = "Jesse David Dinneen, Fabian Odoni"
__copyright__ = "Copyright 2015, JDD"
```

[5] https://github.com/jddinneen/cardinal

[6] Updates to this module – made after the data analysed in Chapter 3 were collected – have improved the list of folders ignored during data collection (e.g., by ignoring the Mac OS Library folder, which was included at the time of data collection in this study). Please see the limitations section of that chapter for a discussion of the significance of the inclusion of the Library folder for the interpretation of the relevant findings.

```
__license__ = "GPL"
__version__ = "0.1"
__maintainer__ = "Jesse David Dinneen"
__email__ = "jesse.dinneen@mail.mcgill.ca"
__status__ = "Beta"


HOME = os.path.expanduser("~")


def scan(locations,  ignores):
    """ Walks through the file system and analyzes the files and folders"""
    startTime = datetime.datetime.now()

    node_lists = []
    temp_filename_set = set()
    temp_foldername_set = set()
    dirs_analyzed = 0
    files_analyzed = 0

    # Search and mark the default folders found in the file system
    if sys.platform in ['Windows',  'win32']:
        default_locations = {os.path.join(HOME, folder) for folder in DEFAULT_FOLDERS["
            win"]}
    elif sys.platform in ['darwin']:
        default_locations = {os.path.join(HOME, folder) for folder in DEFAULT_FOLDERS["
            mac"]}
        apps_folder = [os.path.join(HOME, "Applications")]
        ignores.append(apps_folder)
    elif sys.platform in ['linux',  'linux2',  'linux3']:
        default_locations = {os.path.join(HOME, folder) for folder in DEFAULT_FOLDERS["
            linux"]}
    else:
        raise "Plattform not detected"

    for location in locations:
        norm_location = str(location)
        the_nodes = {}
        temp_nodes = {}
        node_id_counter = 0
        file_id_counter = 0

        def add_depths(node_id, depth):
            """ assigns depth to node, does same for children, their children... """
```

```
    temp_node = the_nodes[node_id]
    temp_node.depth = depth
    for child in temp_node.children:
        add_depths(child, depth + 1)
    the_nodes[node_id] = temp_node

def is_hidden_file(root, f):
    """ checks to see if file is hidden (OS-sensitive) """
    if sys.platform in ['Windows',  'win32']:
        try:
            full_path = os.path.join(root, f)
            attrs = ctypes.windll.kernel32.GetFileAttributesW(full_path)
            assert attrs != -1
            result = bool(attrs & 2)
        except (AttributeError, AssertionError):
            result = False
        return result
    else:
        if str(f).startswith('.'):
            return True
        else:
            return False

def is_symlink_file(root, f):
    """ checks to see if a file is a shortcut or symlink (OS-sensitive) """
    if sys.platform in ['Windows',  'win32']:
        if str(f)[-4:] == '.lnk':
            return True
        else:
            return False
    else:
        if os.path.islink(os.path.join(root, f)):
            return True
        else:
            return False

# the actual walk process and the main observations made at each step
for root, dirs, files in scandir(norm_location,  topdown=True,  onerror=None,
    followlinks=False):
    this_node = Node()
    root_head_tail = os.path.split(root)
    node_id_counter += 1
    this_node.node_id = str(node_id_counter)
    this_node.name_length = len(root_head_tail[1])
```

```
if root in default_locations:
    this_node.default = True

this_node.letters = sum(char.isalpha() for char in root_head_tail[1])
this_node.numbers = sum(char.isdigit() for char in root_head_tail[1])

this_node.white_spaces = sum(char.isspace() for char in root_head_tail[1])
this_node.special_chars = sum(not char.isdigit() and not char.isalpha() and
    not char.isspace() for char in root_head_tail[1])

if root_head_tail[1] in temp_foldername_set:
    this_node.name_duplicate = True
else:
    temp_foldername_set.add(root_head_tail[1])

try:
    root_stat_info = os.stat(root)
    this_node.m_time = datetime.datetime.fromtimestamp(root_stat_info.
        st_mtime).strftime("%Y-%m-%d %H:%M:%S")
    this_node.c_time = datetime.datetime.fromtimestamp(root_stat_info.
        st_ctime).strftime("%Y-%m-%d %H:%M:%S")

    if root_stat_info.st_nlink > 1:
        this_node.hard_link_duplicate = True
except:
    this_node.m_time = -2
    this_node.c_time = -2

original_dirs = len(dirs)
# Removes folders to be ignored from walk
for ignore in ignores:
    norm_ignore = str(ignore)
    dirs[:] = [d for d in dirs if not ((os.path.join(root, d) in norm_ignore
        ) and (norm_ignore in os.path.join(root, d)))]

ignored_children = (original_dirs - len(dirs))
dirs[:] = [d for d in dirs if os.access(os.path.join(root, d), os.W_OK)]

inaccessible_children = (original_dirs - (len(dirs) + ignored_children))
dirs[:] = [d for d in dirs if not d[0] == '.']

this_node.hidden_children = (original_dirs - (len(dirs) + ignored_children +
     inaccessible_children))
```

```
dirs[:] = [d for d in dirs if not (os.path.islink(os.path.join(root, d)))]

# store number of dirs that are actually symlinks
this_node.symlinks = (original_dirs - (len(dirs) + ignored_children +
    inaccessible_children + this_node.hidden_children))

for d in dirs:
    dirs_analyzed += 1
    # print("Scanning directory #{}: {}".format(dirs_analyzed, os.path.join(
        root, d)))
    this_node.path_children.append(os.path.join(root, d))

for f in files:
    files_analyzed += 1
    # print("Scanning file #{}: {}".format(files_analyzed, os.path.join(root
        , f)))

    if is_symlink_file(root, f):
        # increment node.symlinks for each 'file' symlink found (counted in
            the same variable as folder symlinks)
        this_node.symlinks += 1
    elif is_hidden_file(root, f):
        this_node.hidden_files += 1
    else:
        this_file = File()
        file_id_counter += 1
        this_file.file_id = file_id_counter
        this_file.full_name_length = len(str(f))

        try: #sometimes statinfo is not available
            statinfo = os.stat(os.path.join(root, f))
        except:
            statinfo = False

        if statinfo:
            if statinfo.st_nlink > 1:
                this_file.hard_link_duplicate = True

            try: #sometimes even when statinfo is available, particular
                stats are missing
                this_file.a_time = datetime.datetime.fromtimestamp(statinfo.
                    st_atime).strftime("%Y-%m-%d %H:%M:%S")
            except:
                this_file.a_time = -2
```

```
        try:
            this_file.m_time = datetime.datetime.fromtimestamp(statinfo.
                st_mtime).strftime("%Y-%m-%d %H:%M:%S")
        except:
            this_file.m_time = -2
        try:
            this_file.c_time = datetime.datetime.fromtimestamp(statinfo.
                st_ctime).strftime("%Y-%m-%d %H:%M:%S")
        except:
            this_file.c_time = -2

        this_file.file_size = statinfo.st_size
        if this_file.file_size == 0:
            try:
                this_file.file_size = os.path.getsize(os.path.join(root,
                     f))
            except:
                this_file.file_size = -2

    else:
        this_file.a_time = -2
        this_file.m_time = -2
        this_file.c_time = -2
        this_file.file_size = -2

    if str(f) in temp_filename_set:
        this_file.name_duplicate = True
    else:
        temp_filename_set.add(str(f))

    if '.' in f:
        split_name = f.rsplit('.', 1)
        this_file.extension = split_name[1]
        this_file.letters = sum(c.isalpha() for c in split_name[0])
        this_file.numbers = sum(c.isdigit() for c in split_name[0])
        this_file.white_spaces = sum(c.isspace() for c in split_name[0])
        this_file.special_chars = sum(not c.isdigit() and not c.isalpha
            () and not c.isspace() for c in split_name[0])
    else:
        this_file.letters = sum(c.isalpha() for c in f)
        this_file.numbers = sum(c.isdigit() for c in f)
        this_file.white_spaces = sum(c.isspace() for c in f)
        this_file.special_chars = sum(not c.isdigit() and not c.isalpha
            () and not c.isspace() for c in f)
```

```
                this_node.file_list.append(this_file)

        # this is done so that paths need not be permanently stored, but node
            relationships can be recorded
        for temp_id, temp_node in list(temp_nodes.items()):
            if (len(temp_node.path_children) < 1):
                the_nodes[temp_id] = temp_node
                del temp_nodes[temp_id]
            else:
                for path_child in temp_node.path_children:
                    if ((root in path_child) and (path_child in root)):
                        temp_node.children.append(str(this_node.node_id))
                        temp_node.path_children.remove(path_child)

        if len(this_node.path_children) > 0:
            temp_nodes[this_node.node_id] = this_node
        else:
            the_nodes[this_node.node_id] = this_node

    while len(temp_nodes) > 0:
        for temp_id, temp_node in list(temp_nodes.items()):
            while len(temp_node.path_children) > 0:
                for path_child in temp_node.path_children:
                    temp_node.unknown_children += 1
                    temp_node.path_children.remove(path_child)
            the_nodes[temp_id] = temp_node
            del temp_nodes[temp_id]

    node_lists.append(the_nodes)

    add_depths("1", 0)  # call the earlier defined function, starting from the top
runtime = datetime.datetime.now() - startTime
print("Walking runtime: {}".format(runtime))
return node_lists
```