Personal Information Management: From Information Consumption to Curation

Steve Whittaker

IBM Research, Almaden, CA

This paper appeared in B. Cronin (Ed.) Annual Review of Information Science and Technology (ARIST), 45, 1-42, Information Today Inc, Medford, NJ.

<A> Introduction

An implicit, but pervasive view in the information science community is that people are perpetual seekers after new public information, incessantly identifying and consuming new information by browsing the Web and accessing public collections. One aim of this review is to move beyond this *consumer* characterization, which regards information as a *public* resource containing *novel* data that we seek out, consume, and then discard. Instead, I want to focus on a very different view: where *familiar* information is used as a *personal* resource that we *keep, manage*, and (sometimes repeatedly) *exploit*. I call this *information curation*. I first summarize limitations of the consumer perspective. I then review research on three different information curation processes: keeping, management, and exploitation. I describe existing work detailing how each of these processes is applied to different types of personal data: documents, e-mail messages, photos, and Web pages. The research indicates people tend to keep too much information, with the exception of contacts and Web pages. When managing information, strategies that rely on piles as opposed to files provide surprising benefits. And in spite of the emergence of desktop search, exploitation currently remains reliant on manual methods such as navigation. Several new technologies have the potential to address important

curation problems, but implementing these in acceptable ways remains a challenge. I conclude with a summary of outstanding research and technical questions.

<A> Information Seeking and Consumption

There is a long tradition within information and computer science of defining information in terms of its novelty and its ability to transform whoever consumes it (Shannon & Weaver, 1949). Consistent with this, a new set of computer science theories argues that our general information behavior is akin to the foraging behaviors of hunter gatherer peoples (Pirolli, 2007; Pirolli & Card, 1995, 1999). According to this view, people actively *seek out* and *consume* new information from public resources. Although not as extreme in its focus on consumption, the information science literature also emphasizes *discovery* of new public information, rather than its exploitation. Many different models have been proposed to characterize how people find new information in public collections (Belkin, 1980; Ellis & Haugan, 1997; Kuhlthau, 1991; Marchionini, 1995; Wilson, 1981, 1994). As with information foraging, these models focus exclusively on the process of locating new information from public resources (e.g., archives or the Web). Although such information seeking is acknowledged to be iterative, with people making repeated short-term efforts to satisfy information needs (Belkin, 1980; Marchionini, 1995), these models are silent about what happens once such valuable information is located. They do not discuss how this information is organized or curated for future use.

For example, in a very influential theory, Belkin (1980) proposes that people are motivated by ASK (anomalous states of knowledge) to discover new relevant

information. He talks about the steps that people follow to address anomalous states. Kuhlthau (1991) describes various information seeking processes, including recognizing an information need and identifying a general topic, as well as stages for formulating and gathering information. Ellis and Haugan (1997) propose a similar *feature set model* and detail the activities involved in finding information, including browsing, chaining, monitoring, differentiating, extracting, and verifying it. Wilson (1999) provides a high-level macro-model, which characterizes how information needs arise and what aids or hinders these processes of information seeking, incorporating insights from Kuhlthau's and Ellis and Haugan's lower-level accounts. Marchionini's (1995) model is focused on more recent technologies, discussing how information seeking moves from high level framing of information needs to expressing those as some form of query, evaluation of the results of executing that query, and reiteration depending on the outcome of that evaluation.

However, all these models talk about only how public information is *found* and ignore what happens after finding has occurred. In a systematic meta-analysis of theoretical information science, Wilson (1999, p. 4) confirms that information science theories have not tackled what he calls "information use," that is, what happens after information seeking is completed. I will argue that this emphasis on information seeking is based on a partial and unrepresentative view of what people usually *do* with information. In contrast to the foraging and information-seeking viewpoints, this review is concerned with an increasingly important (but very different) set of behaviors which I call *personal information curation*. The collection management literature includes some studies of curation behaviors but these have tended to focus on the activities of

information professionals who are trained to organize and manage *public* collections (Drew & Dewe, 1992; Osburn & Atkinson, 1991). Similar studies within computer-supported cooperative work look at how teams self-organize to create shared repositories (Ackerman 1998; Ackerman & Halverson, 2004; Berlin, Jeffries, O'Day, Paepcke, & Wharton, 1993). In both of these cases, however, the focus of curation is on organization of *public* and not personal collections. Here I review evidence showing that people's everyday information habits are frequently focused around managing personal data and do not involve incessant access and immediate consumption of new public information. Instead, people keep and manage personal information for future exploitation. While reviewing the general literature, I will provide illustrative examples of each of these behaviors from my own research and that of my collaborators.

<B> **Curation Is the Rule and Not the Exception**

One very strong argument for the incompleteness of the consumption model is that people *keep personal* information. Information seeking and foraging models argue that we are continually seeking out novel public resources. If these models are correct, then we should not expect people to conserve large amounts of information for future consumption. However, a minute's reflection will reveal that people persistently engage in active and extensive *preservation* and *curation* behaviors in their information environments. Much as we might want to, we do not immediately delete each e-mail we receive, once we have read or replied to it. And after creating a document or presentation, we do not immediately transfer it to the trash. We take care to preserve personal photos over periods of years.

There are many, many, examples of people preserving and managing personal materials for future exploitation. Here are some simple statistics about the huge amounts of information that people keep in their personal stores. Whittaker, Bellotti, and Gwizdka (2007) summarize eight studies of e-mail use, showing that people archive a huge number of messages, with an average of around 2,846 messages being kept. Unsurprisingly, these personal e-mail archives are growing larger, with more recent studies (Fisher, Brush, Gleave, & Smith, 2006) revealing that people have around 28,000 messages. People also keep a large number of personal files. Boardman and Sasse (2004) found an average of around 2,200 personal files stored on people's hard drives. And a recent study of digital photos found an average of over 4,000 personal pictures (Whittaker, Bergman, & Clough, 2010). Studies of Web bookmarking show that people also preserve hundreds of bookmarks (Abrams, Baecker, & Chignell, 1998; Aula, Jhaveri, & Käki, 2005; Boardman & Sasse, 2004; Catledge & Pitkow, 1995; Cockburn & Greenberg, 2000). And of course these behaviors are not limited to the digital domain: Whittaker and Hirschberg (2001) looked at paper archives and found that people still amassed huge amounts of personal paper data. That study found that on average researchers had 62 kilograms of paper, equivalent to a pile of phone directories 30 meters high.

Furthermore, it is not just that people passively *keep* this information, they also make strenuous attempts to *organize it in ways that will promote future retrieval.* For e-mail, Bellotti, Ducheneaut, Howard, Smith, and Grinter (2005) found that people spend 10 percent of their total time in e-mail filing messages, leading to an average of 244 folders in their e-mail collections. Personal computer files are organized in a similar way, with people averaging 57 folders with an average depth of 3.3 subfolders (Boardman &

Sasse, 2004). Studies of Web bookmarking also show active organizational efforts leading to an average of 17 folders with complex subfolder structure (Abrams et al., 1998; Aula et al., 2005). And Marshall (2008a, 2008b) describes the arcane organizations that result from attempts to preserve information over many years.

So, although it is obvious that consumption is important for some types of rapidly changing transient public information (news, entertainment), it is not the norm. For most types of information, behavior seems to be much closer to *curation* than *consumption*. Furthermore, curation seems destined to become even more important. New technologies—such as ubiquitous sensors, digital video, and digital cameras—make it increasingly easy to capture new types of personal data. And this trend, along with continued increases in cheap digital, mean that people's hard drives are now filling up with huge collections of personal photos, videos, and music (Bell & Gemmell, 2009; Kalnikaitė, Sellen, Whittaker, & Kirk, 2010; Marshall, 2008a, 2008b).

One obvious objection to the argument for curation is that we spend large amounts of time accessing public resources such as the Web. However, new research shows that even here we are not seeking *novel* information. Accessing the Web usually entails *re-accessing previously visited resources*. Various studies have shown that most of people's Web behavior concerns *re-access*, that is, returning to information they have already viewed. Between 58 and 81 percent of all user accesses are of pages that the user has accessed previously (Cockburn & Greenberg, 2000; Obendorf, Weinreich, Herder, & Mayer, 2007; Tauscher & Greenberg, 1997). So, rather than people foraging for *new* information and resources, they instead revisit previously accessed information. Again this suggests a pattern of curation and re-use rather than one-time consumption.

If these arguments are correct, we need to rethink our theories of information. Prior systems and models of information describe consumption of public data. Indeed, until recently it was not possible to create and keep significant personal digital archives. The prevalence of keeping and re-use, however, suggests a need to develop theories of *curation*: the active preservation of personal information content for the future. We need to look beyond models of foraging and information seeking to think about practices of preserving and curating information. Agricultural practices allowed our ancestors to free themselves from the vagaries of an unpredictable environment. In the same way, we need new theories, tools, and practices for information curation to help support these pervasive activities. Although other work has neglected how we acquire and manage personal information, one exception is that of Jones and colleagues (Bruce, Jones, & Dumais, 2004; Jones, 2004, 2007a, 2007b; Jones & Teevan, 2007); we use a variant of Jones's Personal Information Management (PIM) lifecycle framework to organize this review.

The structure of the chapter is as follows. In the next section we present a framework for the *curation lifecycle,* which describes the processes by which we keep, manage, and access information, elaborating the relationships among these processes. We also discuss important distinctions between different properties of information that have implications for curation, such as whether information is unique and whether it requires action. The next three sections review the challenges of keeping, managing, and exploiting personal information. We present relevant research on how and why people keep information, the different ways they organize it, and finally how they access and exploit that stored information. In each case we review how different types of information (e-mail messages, documents, photos, webpages) are treated differently. The

final section looks ahead, exploring different technical developments that may influence the future of information curation, as well as outlining outstanding empirical and methodological issues.

<A> **The Curation Lifecycle**

Curation involves *future oriented* activities, more specifically the set of practices that select, maintain, and manage information in ways that are intended to promote future consumption of that information. We begin by introducing a simple, three-stage model of the curation lifecycle that is a variant of one described by Jones (2007a, 2007b) and Jones and Teevan (2007). We talk about the relations between different phases of the lifecycle and clarify differences between our framework and Jones's work. We also introduce important distinctions between different *properties* of information that have implications for curation behaviors.

<B> **Keeping**

We encounter new information all the time. Much of this encountered information may be irrelevant to us and other pieces of information, such as news or trivia, are of little future utility once we have registered them. But some of this new information we expect to need in the future; how do we decide what is worth keeping? What principles govern decisions about the sorts of information we keep (Jones, 2004, 2007a, 2007b)? There are *costs* to keeping, so how do we decide which information will have significant future value; and what makes it worth keeping (Marshall, 2008a, 2008b)? Keeping is clearly a complex decision that is influenced by many factors, including the *type* of

information being evaluated, *when* we expect we will need it, as well as the *context* in which we imagine that it will be needed. There are also strategic trade-offs involved in keeping information *ourselves* rather than relying on regenerating that same information from public resources.

Information items (whether they are documents, e-mail messages, photos, or webpages) have different utility and will consequently be processed in very different ways. Transient information encountered on a webpage will be treated very differently from a personal document we have been working on for several days or an e-mail sent by an important colleague. The technologies that we use to generate and encounter information will also have an effect on how likely we are to keep it. For example digital photography has now made it much easier to take very many pictures. And preserving digital pictures is inexpensive because storage technology is now so cheap. One consequence is that people are keeping many more pictures, compared with the past when taking pictures was expensive, developing them was laborious, and careful physical organization and storage were needed. But the ease of generating pictures may have important downstream consequences for retrieval that need to be taken into account when deciding whether to keep them (Whittaker et al., 2010).

<B> **Management**

Having decided *that* we want to keep certain information, how should we *manage* that information in ways that will guarantee it will produce future value? Again, this depends on the *type* of information, and once again there are strategic questions. A key decision people have to make is the trade-off between the *effort* to invest in managing

information against the projected *payoff* during exploitation.

The different ways of managing information have different costs and payoffs. As information curators, we have to decide between *intensive* methods that are likely to engender higher information yields but at the cost of greater management efforts. These intensive methods must be compared with less intensive methods that may guarantee less predictable returns. For example, we might apply systematic structure to our paper files by filing incoming information into structured folders. This information should then be easier to access—providing that the structures match the context in which we wish to retrieve the information. However, this filing strategy imposes a heavy burden on the information curator because each new piece of information must be analyzed and structured in this way. Alternatively, we may adopt a more relaxed approach and allow physical information to accumulate in piles on our desk, or e-mail messages to pile up in our inbox. This tactic reduces the costs of organizing the information, but may make it harder to locate critical information when we need it (Malone, 1983; Whittaker, 2005; Whittaker & Hirschberg, 2001; Whittaker & Sidner, 1996).

The management process is also organic and we modify our personal information systems in an adaptive way. We repeatedly revisit and restructure information related to ongoing tasks to meet our current needs. People may be able to remember more about the organization of recently or frequently visited information—making it straightforward to access. Other types of information may be infrequently accessed (e.g., photos that are stored for the long term). Infrequent access may mean that users do not discover that their photo collection needs to be systematically restructured for it to be effectively retrieved (Whittaker et al., 2010).

Management may also have repetitive properties. Some people habitually weed information that has turned out to be of little value and may be compromising the uptake of information with definite utility. People occasionally work through e-mail inboxes, deleting old or irrelevant information (Whittaker & Sidner, 1996). However, it is abundantly clear that people find such cleanup activities difficult, not only because they require judgments about the projected value of information but also because there may be emotional connections with the information that they have invested time and effort in organizing (Bergman, Beyth-Marom, & Nachmias, 2003; Jones, 2007a, 2007b; Marshall, 2008a, 2008b; Whittaker & Hirschberg, 2001).

<B> **Exploitation**

Exploitation is at the heart of curation practices. If we cannot successfully exploit the information we preserved, then keeping decisions and management activity will have been futile. But what are effective ways for accessing curated information? Successful exploitation clearly relates to keeping and management practices. Careful attempts to organize valuable information should make it easier to re-access the data. But technology potentially reduces the need to organize. Emerging technologies, such as desktop search (Cutrell, Dumais, & Teevan, 2006; Dumais, Cutrell, Cadiz, Jancke, Sarin, & Robbins, 2003; Russell & Lawrence, 2007), promise to reduce the overhead of organizing our files because we no longer have to navigate to them manually.

Two main methods can be used to exploit information

<begin bulleted list>

• Navigation—which exploits structures the user has set up for retrieval and involves

incremental manual traversal of these structures.

• Search—a more indirect way to find information—where the user generates textual labels that refer to the name of an information item, one of its attributes, or its contents.

<end bulleted list>

There are advantages and disadvantages to both methods. Navigation, being incremental, offers the user feedback at each access stage (Barreau & Nardi, 1995; Bergman, Beyth-Marom, Nachmias, Gradovitch, & Whittaker, 2008); but in the case of complex folder hierarchies can be laborious because of the multiple levels to traverse. Search is potentially more flexible, allowing users to specify multiple properties of the target file (Lansdale, 1988). However, it is reliant on being able to *remember* salient properties of the target item in order to generate appropriate search terms.

<C> **Relation to Jones's PIM Framework**

The differences in terminology between our framework and that of Jones and Teevan (2007) and Jones (2007a) are shown in Table 1. The frameworks concur in their overall characterization of key personal information management processes, such as keeping and management. However, Jones and Teevan's (2007) main focus is on finding and refinding of public data, for example, if people want to repeatedly access a valued Web resource. In contrast, in this review we are more concerned with information that people create themselves or that they receive in e-mail. Where we focus on Web data it is in the context of users' efforts to integrate such information into their personal information systems. Our stricter definition of personal information means we begin our model with keeping. Keeping is a prerequisite for later stages: People cannot manage or

exploit information that they have not kept. In contrast, Jones and Teevan's (2007) concern is more with public data and they begin with (re)finding such information, because it already exists in the public domain without users making efforts to create or preserve it.

PIM Activities (Jones, 2007; Jones & Teevan, 2007)	**Curation Lifecycle**
(Re)finding	
Keeping	Keeping
Metalevel activities (managing, maintaining, ...)	Management
	Exploitation

Table 1. Comparison of PIM activities proposed by Jones and Teevan (2007) with those used in this chapter

<B> **Interrelations among Keeping, Management, and Exploitation**

As will be obvious from the foregoing, there are close relations among the different processes in the curation lifecycle. For example, successful exploitation is highly dependent on the information people choose to preserve as well as the method they use to manage it. Keeping information does not necessarily guarantee that it will be successfully exploited: The more information we keep, the more effort has to go into organizing and maintaining it. Critically, having more information may increase the difficulty of exploitation because finding what one needs may be harder when there is more information to search.

Past outcomes may also influence future curation behaviors and exploitation success may influence future keeping and management practices. If certain information is difficult to re-access or maintain, people may conclude that there is little point in keeping it in future. In the same way, exploitation failure may cause people to change their

management methods. If users realize that certain types of organization are less successful in promoting access, they may abandon those methods.

<B> **Information Properties**

Not all information items are equivalent. We need to distinguish among various *information properties*, as these differences have implications for how each type of item will be curated.

<C> **Informative versus Action-Oriented Items**

Compare, for example, an average e-mail message, with a page found in a Web search. One crucial property of many e-mail messages is that they require the recipient *to do something*, whether it is to respond to a question, arrange a meeting, or provide some information. Such e-mail messages are *action oriented*, because the message recipient is expected to *respond* in some way, often within a specific timeframe ("let me know about this within the next day"). In contrast, information items found during a Web search are potentially *informative*—but do not usually require users to *act*: The page may be diverting but there is no *requirement* to process the information on the page to meet a given deadline. Of course this information vs. action distinction does not map neatly to computer applications. Not every message in e-mail is action oriented (e.g., when people send us FYIs) and not every webpage is purely informative (e.g., when it contains a request to complete a form).

This distinction has critical implications for how we treat personal information. For reasons that will become clear, it is often impossible to discharge *action-oriented*

items immediately. So, *reminding strategies* (e.g., creating task related e-mail folders or leaving active documents on the desktop) have to be set up to prompt users about their commitments with respect to the undischarged item. Failure to set up such structures can have severe implications for job success and productivity; we must not forget to respond to that important request from our boss, even when we are inundated with other commitments. In contrast, how we deal with *informative* items is usually more discretionary: They usually do not need to be actively processed to meet deadlines, so it is less critical that we create dedicated reminding structures to ensure that they are dealt with appropriately.

<C> **Information Uniqueness**

Another critical information property is *uniqueness.* Uniqueness has strong implications for how we deal with personal information. Certain types of information (such as personal files that we create ourselves) may be resident *only on our computer*. As a result we may be the only person in the world who has access to those items. Anyone who has lost data following a system crash is only too aware that if we do not take responsibility for storing and maintaining unique data, it will not be preserved for future access (Marshall, 2008a, 2008b). In contrast, public information such as Web data may be resident on multiple servers and may be recoverable even if we personally take no action to store a local copy. E-mail data lie somewhere in between. We may be able to ask coworkers to regenerate a copy of an important message that we have temporarily mislaid, or lost in a system crash, but we cannot guarantee they will have kept that information.

It is important to note that uniqueness is defined *subjectively*: relative to our own goals and interests. Innumerable unique information items exist in the world, but as curators we are concerned only to take decisive action to preserve those that are *relevant to us*. Other people's information may be equally important to them, but there is no reason we should be concerned to preserve it, unless of course we work with them. This personal uniqueness is often associated with information that we have invested *effort* in generating. If we have dedicated substantial time to generating an information item (e.g., an extended personal document, a carefully crafted presentation, or a collection of wedding photos), then that information will be something that we make enormous efforts to preserve, in part because of the effort involved in regenerating it.

Uniqueness has a huge impact on our management strategies. None else will take care of our unique personal data. We personally need to create reliable structures for re-accessing highly personal data such as passwords, tax forms, passport details, or financial records, even when we rarely need to access this information. Personally produced documents also tend to be unique and need to be carefully organized. The same is largely true for e-mail messages: We need to have reliable methods for re-accessing these because we cannot always rely on others to keep the important messages that we need. Webpages are rather different. They are generally more easily recoverable (via search or browsing) even if we have not bookmarked them. And in addition, because we have not usually been responsible for generating their content, we are not as concerned if we cannot recover the information they contain.

Table 2 shows the key properties of common classes of information, such as paper, electronic files, e-mail, photos, and Web documents. The table shows these are

very different with respect to action orientation and also uniqueness. Current paper and electronic documents and e-mail messages are often action oriented. Paper and electronic documents and personal photos tend to be unique. These differences have strong implications for curation. The uniqueness of paper or electronic personal documents and personal photos leads people to be very conservative and to keep most of these items. They also have to preserve action oriented items such as e-mail messages and personal documents in ways that promote effective action.

Information Type	**Action or Information Oriented?**	**Uniqueness**
Personal paper documents	Action oriented if self created and current Long term archives tend to be informational	Unique if self-created or annotated
Personal electronic documents	Action oriented if self created and current Long term archives tend to be informational	Unique if self-created
E-mail	Often action oriented Long term archives tend to be informational	Range from unique to non-unique mass mailings
Personal photos	Affective	Predominantly unique
Web	Informational	Non-unique

Table 2: Main properties of different information types

We now turn to each of the main curation processes, describing how people keep, manage, and exploit their personal information. It will be clear from our prior discussion that there are huge dependencies among these processes. In what follows, we review each of these processes separately; but we should not lose sight of the relationships among them.

<A> **Keeping**

<B> **Overview, Problems, and Strategies**

We encounter too much information to keep it all because of various costs including:

<begin bulleted list>

• Management costs—we need to organize information if we are to obtain value from it. The more we keep, the more management effort is required. Some visions of new technology suggest that in the future our information will be organized automatically, but these technologies are not yet in place. Indeed, in future sections, we explore whether these technologies will *ever* effectively replace the need for manual organization.

• Exploitation costs—keeping information of low value increases the difficulty of retrieval. Keeping too many items can be distracting if manual browsing is used for access. Nevertheless, there are those who contend that future retrieval will be entirely *search based*, reducing exploitation costs regardless of how much we keep.

<end bulleted list>

Keeping decisions are a fact of life. Every day we receive new e-mail messages, create new files and folders, and browse new websites. Some of this information is of little long-term value but some of it is task-critical and needs to be preserved for the long term. Boardman and Sasse's (2004) data suggest that users acquire an average of five new files per day, five e-mail messages[1] per day, and one bookmark every five days. Other studies indicate people acquire one new contact per day (Whittaker, Jones, & Terveen, 2002a), and around five digital photos (Whittaker et al., 2010). But these

statistics are an over-simplification of the complexity of keeping decisions. The statistics record *positive* decisions but fail to register the many decisions to reject information judged to be of little value. To be more specific, we know that users receive an average of 42 e-mail messages per day; so focusing exclusively on the five they actively decide to keep overlooks the 37 decisions they make to delete irrelevant information. For e-mail alone, making the highly conservative assumption that e-mail volumes will not change over our lifetimes, this equates to around 1 million keeping decisions over a 60-year digital life.

We know from various interview and survey studies how difficult people find it to decide what information they want to keep (Bergman, Tucker, Beyth-Marom, Cutrell, & Whittaker, 2009; Boardman & Sasse, 2004; Jones, 2004, 2007a, 2007b; Whittaker & Hirschberg, 2001; Whittaker & Sidner, 1996). But why are keeping decisions so difficult? One reason is that they require us to predict the future. To decide what to keep, we have to determine the probable future value of an information item.

This may be a general psychological problem. A great deal of psychological research shows that people are poor at making many types of decisions that involve their future. Such prediction requires people to reason about hypothetical situations, at which they are notoriously poor. People's predictions are also subject to various types of bias. For example, they expect the future to be very much like the present; and their predictions are unduly influenced by recent, or easily recalled, events (Gilbert, 2006; Kahneman & Tversky, 1979).

In the PIM context, the keeping decision requires people to predict future informational contexts and assess future informational requirements. Jones (2004, online)

argues that the decision whether "to keep or not to keep" information for future usage is prone to two types of costly mistakes. On one hand, information not kept is unavailable when it is needed later. On the other, keeping irrelevant information not only causes guilt about being disorganized (Boardman & Sasse, 2004; Whittaker & Sidner, 1996), it also increases retrieval time. Irrelevant information competes for the user's attention, obscuring important information relevant to the current task. Indeed, it is well known in psychology that in visual search the number of irrelevant distracters increases the time taken for people to identify a target object (Treisman & Gelade, 1980). Furthermore, there is a "deletion paradox": although unimportant information items distract attention and increase retrieval time for important items, it takes time and effort to review items to decide whether to delete them (Bergman et al., 2009, p. 1).

When people weigh the advantages of keeping versus deleting, some of the reasons for keeping are rational—after all, the user can always think of a situation when the information item may be needed (Whittaker & Hirschberg, 2001). However, there are also less rational reasons why people avoid deletion, which can be attributed to general psychological decision-making processes (Kahneman & Tversky, 1979). In making decisions, losses and gains are evaluated *asymmetrically*: Losses are more salient than gains and the possible loss of information emotionally influences the decision maker more than the gains of reduced retrieval time. Small objective probabilities are subjectively weighted more highly than their actual likelihood. Thus people perceive as significant the very small probability that a deleted information item will be needed.

We now review various studies looking at keeping decisions people make, when managing their paper archives, e-mail messages, contacts, webpages, and personal

photos.

<B> **Keeping Paper**

Somewhat curiously, in spite of the prevalence of keeping decisions, relatively few studies have looked at this directly. One exception is a study of people's paper archiving behavior (Whittaker & Hirschberg, 2001). Although there is a common intuition that the world is shifting away from paper and becoming more digital, we will see that people treat paper in ways that are very similar to their treatment of digital information.

One methodological problem with investigating keeping behavior lies in finding contexts where people are explicitly focused on the keeping decision. Our study identified one such situation. Participants were about to move offices and had to decide about which information to keep and what to throw away. When we interviewed them, they had all recently sorted through their paper archives in preparation for the move. Their new offices had reduced personal storage space compared with their existing offices, although extra storage was provided in public locations. This reduction in local storage motivated careful reflection as well as sorting and discarding existing data. In interviewing and surveying workers, we capitalized on the fact that they had very recently handled most of their paper data, forcing them to identify criteria for determining what to keep and what to discard.

<C> **Discarding Behavior**

People experienced major problems in deciding what to keep and what to throw

away. As the psychology literature would suggest, there was a bias toward preservation. Even after spending large amounts of time deciding what to discard, workers still retained huge archives after the move. In preparation for the move, people spent almost nine hours rationalizing their data and reported that this was a difficult process. In spite of these efforts, the amount of information that people actually threw away was small compared with what they kept: They discarded just 22 percent of their original archives, with the final preserved archive averaging more than 18 mover's boxes.

We looked at the characteristics not only of what people kept, but also of what was discarded. As expected, at least some of the discarded data were once-valuable information that had become *obsolete*. As jobs, personal interests, or company strategy changes, the value of particular information decreases. But not all discarded information underwent the transition from valuable to obsolete. For example, 23 percent of discarded data were *unread*. Two general problems led to this accumulation of superfluous information. First, people experience problems with *information overload*, leading them to only partially process incoming information. Second they engage in *deferred evaluation* of what to keep—causing them to acquire large amounts of data that later turn out to be extraneous.

Information overload occurs when people have insufficient time to process all the information to which they are exposed. One consequence of information overload is that non-urgent information is never processed. Non-urgent data are set aside (often in optimistically named "to read" piles), accumulating indefinitely, because the same time pressures that prevent complete processing of incoming data also prevent rationalizing ("clean-up") of archives. Consequently, people seldom discover that their unread non-

urgent documents are superfluous until exceptional circumstances (such as an office move) force them to scrutinize their archives.

Yet even when people find the time to examine new information systematically, uncertainty of the future value of that information means they are often highly conservative—postponing final judgments about utility until some unspecified future date. Some people deliberately *defer evaluation* about incoming information, allowing time to pass so as to make better informed judgments about information utility. Often these post-hoc judgments are based on whether information was ever actually used.

Deferred evaluation means people retain information of unclear value—*just in case* it later turns out to be useful. Finally, judgments about potential utility are made more difficult because the value of data can change over time. Knowing that the value of information might change also leads some people to postpone the keeping decision while there is still archival space.

Accumulating unprocessed data and deferring evaluation are good from the (conservative) perspective that potentially valuable information is not lost. However, the problem with this approach is that people seldom revisit their archives to rationalize them, so they end up containing considerable amounts of information of dubious value. Thus, 74 percent of our users had not cleaned out their archives for over a year. Furthermore, very few clean-ups occur spontaneously: 84 percent arise from extrinsic events such as job changes or office moves. This infrequency of clean-ups means that documents are often not discovered to be superfluous until they have been stored for some time.

To sum up, our deletion data illustrate important aspects of keeping. When

extraordinary events such an office move occur, people discard about 22 percent of their data, some of which is obsolete. However other factors beside obsolescence, such as information overload and deferred evaluation, mean that archives are polluted by marginally relevant data. Rather than discarding once-valuable information that is now of little utility, much of what people later discard is unprocessed information they have never properly evaluated, or kept just in case.

<C> **What Do We Keep and Why Do We Keep It?**

We also looked at the properties of the information people kept and their reasons for keeping it. One conjecture was that a large proportion of the information kept would be unique to that person. In contrast, we expected people to be much less likely to keep publicly available data. Why take responsibility for data that are available elsewhere?

Uniqueness was clearly important in determining whether users would preserve certain documents. Unique data are usually highly associated with their archiver. Three types of unique data accounted for 49 percent of people's archives: working notes, archives of completed projects, and legal documents.

But contrary to our expectations, uniqueness was not the sole criterion for deciding to keep data. Only 49 percent of people's original archive was unique: 15 percent was unread, but 36 percent consisted of *copies of publicly available documents*. We have already discussed why people preserve unread data, but why keep duplicates of public documents? Four main reasons were given: availability, reminding, lack of trust in external stores, and sentiment.

Availability allows relevant materials to be at hand when people need them.

Several people mentioned not wanting to experience the delay associated with refinding information, or even accessing it on the Web. In other words they wanted to reduce their exploitation costs by keeping valued information in a personal archive.

Reminding relates to availability. A personal copy prompts people about outstanding actions associated with a document, or simply reminds them they are in possession of that information. Documents in public or digital stores seemed less capable of supporting reminding. People also kept personal copies of public data because they did not trust other archival institutions to keep the documents they needed. Distrust of external stores also extended to digital resources such as the Web.

In addition to these functional reasons, people described sentimental reasons for keeping information. People admit such information has little relevance for likely future activities but they still cannot part with it because it is part of their intellectual history and professional identity.

Another potential reason for keeping personal copies of publicly available documents is that they contain *personal annotations*. Other research has documented the utility of annotations for focusing attention and improving comprehension of what is read or heard (Kalnikaitė & Whittaker, 2007, 2008a; Sellen & Harper, 2002). Although most people made such annotations, they seemed of little long-term use. Many people stated that annotations had transient value, becoming uninterpretable after some time has elapsed. This is consistent with recent studies of long-term note taking, showing that the utility of handwritten notes decreases even after a month (Kalnikaitė & Whittaker, 2007, 2008a).

<B> **Keeping E-Mail**

E-mail is different from either self-created files or documents accessed on the Web. One major difference is that a significant proportion of the information we receive in e-mail is *actionable*: we have to respond to it or process it, often within a specific time frame. This contrasts with most Web-based information which does not demand action. Another significant difference is that most e-mail messages are generated by others (who in some cases are unfamiliar to the main user). This lack of familiarity sometimes makes it harder for people to decide on the utility of such e-mail information. A final, rather different, characteristic of e-mail is its sheer variability. In our inboxes we may see many different types of messages, including: tasks or to-do items, documents/attachments, fyis, schedules, social messages, and jokes. Again, this heterogeneity makes the keeping decision rather different from that for other information types.

Overall we keep about 70 percent of our e-mail messages (Dabbish, Kraut, Fussell, & Kiesler, 2005). This seems a surprisingly high retention rate, given the apparent irrelevance of many of the e-mail messages we receive, but there are reasons for this. We shall separately discuss people's keeping behaviors for informational versus actionable messages, because the keeping behavior is very different.

<C> **Informational Messages**

Informational messages form about one third (34 percent) of what is delivered in e-mail (Dabbish et al., 2005). Informational messages are treated in a similar manner to paper documents. As with paper, the keeping decision is often difficult. People find it hard to judge the value of incoming informational messages, so they use the *deferral*

strategy. Rather than investing valuable time in reading a new informational message, users register its arrival but defer dealing with it until they are more certain of its value. Deferred e-mail messages are kept around, allowing more informed judgments to be made later.

Users are aware that deferred messages need to be re-evaluated at a later point. Some employ folders for this purpose: 28 percent of informational messages are filed for later reading (Dabbish et al., 2005). However, the problem with this strategy is that filing may lead these messages to be out of sight and out of mind because these folders are seldom revisited (Whittaker & Sidner, 1996). A more common strategy is to leave them in the inbox: Dabbish and colleagues found 42 percent of informational messages are left in the inbox, increasingly the probability that deferred evaluation will actually take place. The inbox is an active workspace: Leaving information there increases the chance that information will be revisited as users re-access the inbox to process incoming messages. But there is an obvious downside to this strategy. Although the strategy increases the probability of revisiting yet-to-be-decided items, the presence of such unevaluated information makes it more difficult for people to locate important information, such as messages requiring action (Bellotti, Ducheneaut, Howard, & Smith, 2003; Whittaker, 2005; Whittaker & Sidner, 1996).

As with paper archives, people experience information overload in e-mail. Overload may lead people to defer completely reading each message until they have more time. And of course because they are constantly bombarded with more incoming messages, they often never return to deferred messages (Whittaker & Sidner, 1996). One factor contributing to whether a message is read is its length; Whittaker and Sidner found

that 21 percent of inbox messages contained enough text to fill more than five screens, consistent with the fact that people leave longer messages there for later reading.

<C> **Actionable Items**

Actionable messages are those with which we have to do something specific. In an ideal world (such as that inhabited by management consultants), we might process these messages just once, carrying out the required action and then deleting them. This is often referred to as the one-touch model. The advantages of the model are obvious: Touching a message just once means that users do not forget to deal with it; and they do not have to repeatedly reconstruct the context of old messages when they eventually come to process these. And if messages are processed at once this keeps the inbox clear for important incoming messages.

Some users try to adhere to this model: Overall, users reply to 65 percent of actionable messages immediately (Dabbish et al., 2005). An immediate reply clearly reduces the chance that one will forget to act on a message. However, even when people do reply immediately *they still keep 85 percent of actionable messages*, suggesting that one touch does not describe actual practice.

Several reasons may account for such retention. In some cases, one touch and an immediate reply are not possible. Many important e-mail tasks are too complex or lengthy to be executed immediately (Bellotti et al., 2005; Venolia, Gupta, Cadiz, & Dabbish, 2001; Whittaker, 2005; Whittaker & Sidner, 1996). This leads to deferral of 37 percent of actionable messages (Dabbish et al., 2005). Deferral is often a direct consequence of *interdependent* tasks: those involving tight collaboration with others

(Bellotti et al., 2005; Whittaker, 2005). Interdependence results in both iteration and delays between messages relating to the task. Iteration arises because interdependent tasks often require multiple exchanges between participants (Bellotti et al., 2005; Venolia et al., 2001; Venolia & Neustaedter, 2003; Whittaker & Sidner, 1996). People may need to negotiate exactly what a collaborative e-mail task involves or who will be responsible for each component. This consensus needs to be built and multiple responses often need to be collated. Delays occur because these negotiations take time and because collaborators often lack the necessary information to respond immediately to address their parts of the task. One way to estimate the prevalence of interdependent tasks is by determining how many e-mail messages are part of a conversational thread, because threads indicate relations and common underlying activities among messages. Threading estimates range from 30 to 62 percent of messages (Bellotti et al., 2003; Whittaker et al., 2007).

The need to defer actionable messages has important consequences for keeping. Unless actions are discharged, messages are usually kept around as reminders that they are still incomplete. Actionable messages are therefore almost always kept (only 0.5 percent are deleted). This figure is much higher than for information messages, 30 percent of which are deleted. Furthermore, actionable messages have to be kept in a way that guarantees that they will be reencountered. It is no good deferring to-do e-mail messages unless you have some method of guaranteeing that you actually return to them. We revisit this issue in the next section, when we talk about management strategies.

<B> **Keeping Contacts**

Contact management is another area that demands careful keeping decisions. Whittaker, Jones, and Terveen (2002a) looked at people's address books, rolodexes, calendars, and contact management programs to explore the criteria they used for including someone in their contact lists. We are overloaded with respect to the contacts we encounter. We are copied on many messages and we read webpages or blogs from friends, colleagues, and strangers. Some of these are people with whom we want to interact again. Others may have been involved in one-off conversations that require no follow-up. Contact management requires decisions as to the people about whom you will keep contact information, as well as the types of information that you will keep about those people.

Deciding on important contacts from the many people that you are exposed to on a daily basis is complex. As with paper and e-mail archives, it is hard to know whether you will need to communicate with that person in the future: whether someone is an *important contact* becomes clear only over time. Just as with the deferral strategy, our informants often *over-saved* information, leading to huge rolodexes, overflowing booklets of business cards, and faded post-it notes scattered around their work areas. But in spite of this strategy, participants were exposed to many more contacts than those about whom they recorded information.

We identified specific factors that were critical in determining important contacts. Just as with deferred evaluation in e-mail and paper archives, the final decision to keep depends on past interaction with the contact, in particular *frequency* and *recency* of communication. People also noted how difficult it was to make decisions about the future based on short term interactions and scanty evidence. Again we see the importance of

long term information in evaluating contacts: important contacts are those with whom we have repeated interactions over extended periods. In addition, the selection process is error-prone because of the difficulty of predicting long-term relationships on the basis of brief initial interactions.

In a follow up study, we presented people with contacts mined from their e-mail archives and asked them to distinguish between important and unimportant ones. The findings were striking. In spite of having huge archives of contacts (858 on average), participants rated only 14 percent (118) as important and worth keeping. Criteria for inclusion echoed those identified in our earlier interviews: Participants chose contacts with whom they interacted frequently and recently, as well as for a long time, and who were likely to respond to their e-mail messages. They excluded spammers.

Overall, interesting parallels appear among contacts, paper, and e-mail messages. People are exposed to many more contacts than they can record systematic information for, so they reserve judgment and overkeep data about contacts they do not need. Furthermore, the criteria people use to judge the value of contacts are based around usage and interaction: Valued contacts are those with whom they interact frequently and recently. However, one key difference between contacts, e-mail, and paper is that users ignore or discard a much higher percentage of encountered contacts.

<B> **Keeping Web Pages**

Similar problematic keeping decisions also surface on the Web (Jones, 2004), where we see errors of commission (over-keeping information that turns out to have little future value) and omission (failing to keep information that turns out to be needed later).

There are clear errors of commission; for example, people expend energy creating bookmarks that they never subsequently use. Tauscher and Greenberg (1997) showed that 58 percent of bookmarks are never used, suggesting poor decision making.

At the same time, other studies of Web behaviors reveal failures of omission—where people do not preserve information that turns out to be useful later. Wen (1993) coined the term *post retrieval value* to describe Web resources that people have accessed but not preserved—only later realizing their utility. His study showed that people were able to later find only about 20 percent of information they have previously accessed and attended to in an earlier information retrieval session. Such failure originates in part from an unwillingness to make deliberate attempts to keep information; his users were unwilling to create bookmarks as records of useful pages because these would clutter their current bookmark collection. These findings were replicated in similar studies (Aula et al., 2005). Instead, users preferred to try to retrace their original searches—a strategy that is often unsuccessful.

<B> **Keeping Photos**

With the advent of digital photography, the number of pictures that people are now taking has increased massively (Bentley, Metcalf, & Harboe, 2006; Kirk, Sellen, Rother, & Wood, 2007; Whittaker et al., 2010, Wilhelm, Takhteyev, Sarvas, Van House, & Davis, 2004) and similar keeping issues are beginning to arise for digital photos.

We looked at this in a study of parents with young families (Whittaker et al., 2010) who had an average of 4,475 digital pictures. All participants deleted some pictures, both when pictures were taken and when they were uploading from camera to

computer. Participants estimated they deleted on average 17 percent of their pictures. The reasons they gave for deletion were that the pictures were of poor technical quality or did not capture an event of interest. In general, deletion was a difficult process, as evidenced by the fact that many of the pictures that were kept were near duplicates (i.e., multiple pictures of identical scenes), an observation that is confirmed in other studies (Kirk et al., 2006), suggesting that people are keeping their options open about the best view of a given scene. One of the reasons people gave for this overkeeping was that they perceived little cost in keeping many photos. They were not, therefore, focused on the exploitation/retrieval context when they made keeping decisions. As with paper and e-mail, people had a strong expectation that they would return to their photo collection to rationalize it at a later date. And as in our paper and e-mail studies, this rationalization seldom occurred.

<B> **Keeping Summary**

Keeping decisions are difficult because they require people to: (1) predict their future retrieval needs, (2) take into account the possibility that those information needs may change, and (3) make utility decisions under conditions of information overload, often on incomplete readings of information.

Errors are made: the primary tendency is overkeeping—keeping things that are never accessed (observed with paper, e-mail, contacts, and photo archives)—although there is evidence from some Web studies of failing to keep information that later turns out to be relevant.

Consistent with overkeeping, deletion is relatively infrequent, varying from 17

percent for photos to 30 percent for e-mail messages. Contacts are very different, however; it seems that because people are exposed to many more of these, they are happy to ignore 86 percent of the contacts they encounter.

The nature of the information item affects the keeping decision. This decision is relatively straightforward for certain items: We obviously need to keep actionable e-mail messages that have not been handled or unique personally generated items that no one else will safeguard. However, it is very hard for people to determine the value of data such as public Web pages or informational e-mail messages.

Rather than viewing keeping as a one-time decision, people often used a deferral strategy—waiting to see whether information was useful. Two major weaknesses of deferral are: (1) people seldom return to their collections to carry out a re-evaluation of tentatively kept information; (2) deferral means that collections are full of items of dubious value, which makes it more difficult to find truly valuable information.

People do not generally seem to be aware of the implications of overkeeping. Although they complain about how full their inboxes are, they nevertheless delete only 30 percent of e-mail messages; even after spending days working though paper archives they still preserve 78 percent of those. On the Web, in contrast, there is a suggestion that people do not bookmark because they are aware that this will make valued materials harder to find. This could be because they consider Web information to be unimportant or because they think it is easily recoverable by other means.

<A> **Management**

<B> **Overview, Problems, and Strategies**

We first describe different methods for organizing information, as well as the trade-offs among them. We next discuss factors that influence users' choice of management strategies and studies evaluating the utility of these strategies. We then briefly talk about a radical alternative, which proposes that we forgo preparatory organization altogether and rely totally on search for information exploitation.

Management is a crucial curation process because it directly affects exploitation. We are constantly acquiring information and, over long periods, large amounts of personal information accumulate (Marshall, 2008a, 2008b). Using current estimates of how many documents, digital photos, and e-mail messages we acquire on a daily basis (Boardman & Sasse, 2004; Whittaker et al., 2010), and making the conservative estimate that these will remain constant over our digital lifetimes, we will actively save around 125,000 documents, 115,000 e-mail messages, and 120,000 digital photographs.

Certain types of management also take place more often than we might expect. For certain items, such as files and e-mail messages, people are perpetually and actively engaged in re-organization, as reflected by the frequent small modifications they make to their information. For example, a longitudinal study (Boardman & Sasse, 2004) found that people create a new file folder every three days and they make a new e-mail folder every five days. In each case, the new structure demonstrates that people are constantly reflecting on how their information is currently organized and finding it to be inadequate. However, as mentioned in the keeping section, people seldom engage in major reorganizations or extensive deletion. Instead, they tend to modify existing structures incrementally. They are highly unlikely, however, to monitor and re-organize photos or

contacts, for reasons that will become clear.

People also make management mistakes. They often engage in counterproductive behaviors in organizing their information. Studies of Web bookmarking show that people construct complex hierarchical bookmarking systems (Abrams et al., 1998; Aula et al., 2005). Yet, we have already seen that users *never* access 42 percent of the bookmarks they organize for later retrieval (Tauscher & Greenberg, 1997). Efforts at organizing e-mail messages may also not bear fruit. E-mail filing accounts for 10 percent of total time in e-mail (Bellotti et al., 2005), yet information is usually accessed by browsing the inbox or search, rather than folder access (Tang, Wilcox, Cerruti, Badenes, Nusser, & Schoudt, 2008; Whittaker, 2005; Whittaker et al., 2007). With personal photos they may make the opposite type of mistake and fail to organize information when there is a clear need to do so. For example, a study of personal photo retrieval showed a failure to impose even rudimentary organization—in part because people believed that they would be able to retrieve their photos without needing to organize them (Whittaker et al., 2010).

<C> **Semantic Organization**

Organizing information is a fundamental cognitive activity. One basic approach is to apply conceptual organization to information. Even newborn infants categorize objects, with natural psychological categories tending to be based around exemplars or prototypes. For example, an individual's concept of bird is based around exemplars such as robins, rather than unusual cases such as penguins. Our judgments and reasoning are influenced by the extent to which particular instances are similar to those exemplars (Rosch, 1978; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

When managing personal information, two different and separate aspects to organization are important for effective exploitation. We call these *mental* and *external* cueing. As many psychological studies have shown, the mental act of imposing organization on information makes it inherently more memorable. Organizing things within a consistent conceptual structure means that, at recall, one item may trigger memory of a related one; therefore, applying semantic organization is highly effective in promoting recall (Baddeley, 1997; Craik & Lockhart, 1972). Organization helps recall, even if people do not have direct access to their organizational scheme at retrieval. For example, in a recent study we showed that the simple act of organizing conversational information by taking notes increased recall even when people did not use their notes at retrieval time (Kalnikaitė & Whittaker, 2008b). Organization is also important because the products of organizational efforts can themselves be used as external retrieval cues. Appropriate notes can serve as cues to remind us about information items that we might otherwise have forgotten (Kalnikaitė & Whittaker, 2007, 2008b). Well chosen folder names cue people about their contents and organization (Bergman et al., 2003; Jones & Dumais, 1986; Jones, Phuwanartnurak, Gill, & Bruce, 2005; Lansdale, 1988).

Organization and labelling are mainstays of most computer operating systems. The primary way people organize their digital information is to sort it recursively into categories (in directories, folders, or subfolders) and then apply meaningful labels to these folders and subfolders. The act of applying organization may help retrieval by mental cueing, as well as generating a navigable conceptual structure with folder labels serving as external retrieval cues. Note also that folders usually contain a strong spatial component—with subfolders sitting inside; this, too, can help cue retrieval (Jones &

Dumais, 1986).

<C> **Temporal Organization and Reminding**

A second, less obvious, type of organization has been less extensively researched. We have already seen that some important information with which people deal is *actionable.* Further, it is usually the case that those actions are *required to happen by a certain time*, for example, to meet a certain deadline. People must therefore ensure that actionable information is organized in such a way that it is encountered at the right time*,* allowing the deadline to be met. This is the problem of *reminding*. It is no good having an extensive organizational structure allowing access to any item, if you forget the deadline relating to that information. Reminding is a critical problem, especially in the case of e-mail, where actionable items are prevalent.

Most psychology research on organization has looked at natural categories (e.g., how we mentally organize places, events, names, and faces). It has not looked at the types of information we are addressing here, namely *synthetic,* human-generated information such as documents, e-mail messages, photos or webpages. Nevertheless human-computer interaction (HCI) considerable HCI and library science research has looked into people's preferences for organizing such personal data. For example, people prefer to relocate their documents *spatially* rather than using keyword search (Barreau & Nardi 1995; Bergman et al., 2008). This spatial organization works even better when the document space is three-dimensional, although this may not scale well to large numbers of files (Robertson, Czerwinski, Larson, Robbins, Thiel, & van Dantzich, 1998). However, there are limits to the utility of spatial organization: Semantic labels are

stronger retrieval cues than spatial organization alone, although combinations of semantic and spatial organization can enhance performance (Jones & Dumais, 1986). And semantic and spatial cues are enhanced when these are *self-selected*, rather than being chosen by an external party (Bergman et al., 2003; Lansdale & Edmonds, 1992). There is also evidence for the utility of temporal organization as a retrieval cue. People can successfully retrieve documents by associating them with personal or public events that happened close to the time that the documents were encountered or created (Ringel, Cutrell, Dumais, & Horvitz, 2003). The importance of temporal factors is also shown by log files of search tools revealing a bias toward retrieval of *highly recent* information (Cutrell, Dumais, et al., 2006; Dumais et al., 2003).

In addition to these overall organizational preferences, other work has explored different types of management strategies and what motivates people to choose them. We now describe strategies for paper, digital files, e-mail, Web documents, and photos. We review the types of management strategies employed, what influences people's choice of strategy, and the trade-offs between strategies.

Several recent papers have argued that manual organization of our personal data will soon become obsolete. Improvements in desktop search will mean that documents, e-mail messages, and webpages can be easily retrieved without the need for active organization (Cutrell, Dumais, et al., 2006; Russell & Lawrence, 2007). This is an appealing idea. We have seen that management activities are onerous and difficult for users, who may invest in organizational efforts that are not always directly successful. We consider these claims in more detail when we discuss exploitation techniques and evaluate the efficacy of different search tools.

<B> **Managing Paper**

Malone (1983) conducted a pioneering study into people's organizational habits for paper, identifying two main strategies he called filing and piling. *Filing* involves constructing an exhaustive, hierarchical taxonomy, with semantically related items stored in each subcategory. In contrast, *piling* is more laissez-faire, usually resulting in shallower, less systematic hierarchies. Piles tend to be fewer in number with each pile containing more items, with looser associations between items stored in the same pile. Items may also be in a common pile because they were first generated or acquired at the same time.

There are clear trade-offs between these two organizational strategies. Piles are easier to create and maintain because they are less systematic. They have a less clear organizational structure with more items in each pile, which may make retrieval within each pile less efficient. But because there tend to be fewer piles in total, this leaves fewer potential locations to be searched, which may compensate for the lack of organization. Fewer piles may also mean that users visit each pile more frequently and end up being more familiar with the contents of each. Files, in contrast, require more effort at creation time and more maintenance. However, they offer benefits at retrieval, providing a more coherent retrieval structure along with more relevant labels as cues. These advantages may be offset by the fact that there may be more categories, so files may have more levels to navigate. Files may also fall into disrepair, with too many levels/distinctions being too infrequently visited, making distinctions between categories harder to remember.

In the move study described previously, we investigated when and why people choose filing or piling strategies. The distinction between filers and pilers was not absolute, being instead one of degree. All our respondents filed some information but kept other information in desktop piles. We classified users according to how likely they were to file information. Based on the predominant strategies that people described in our interviews, we identified a threshold of 40 percent for people to be categorized as filers.

Pilers often amass information without attempting to organize it systematically. This laissez-faire approach should lead to an accumulation of unscrutinized information before the office move. We found to our surprise that pilers had smaller original archives. They also had less preserved information than filers after cleaning out their archives. Why then did filers amass more information? Our interviews suggested one possible reason is *premature filing*: filers may file information that turns out to be of little utility and must be discarded later. If filers are more likely to incorporate documents of uncertain quality into their filing systems, we might expect them to throw away more reference materials than pilers in preparing for the move. This was not true for all documents, but was true for reference documents.

There were also differences between strategies in terms of data acquisition. We expected pilers to acquire information faster because they tend not to scrutinize incoming data as carefully. We looked at data acquisition rates, in separate analyses of original and preserved (i.e., post-move) information volumes. For both measures, pilers tended to be slower to acquire original as well as preserved information, when we allow for the number of years they had been in the company.

Given their more systematically organized systems, we expected filers to have an

easier time of finding data and that they would access their data more often. Contrary to our expectations, pilers had accessed a greater percentage of documents than filers in the last year. Why were pilers more likely to access recent data? The interviews revealed both strategies had strengths and weaknesses. With a piling strategy, information is more accessible: It can be located in a relatively small number of piles through which people frequently sift. The result of this is that valuable, frequently accessed information moves to the top of the piles and less relevant material ends up located lower down. This pattern of repeated access allows people to identify important information, discarding unused or irrelevant information.

But the lack of a coherent system with piling has some disadvantages. Taken to excess, piles can dominate not just working surfaces but all areas of the office. However, even though filing is more systematic, it does not always guarantee easy access to information. With complex data, filing systems can become so arcane that people forget the categories they have already created, leading to duplicate categories. Accessing only one of these duplicates leads to incomplete retrieval because some part of the original information will be neglected. This illustrates a general disadvantage to filing strategies: They incur a large overhead for constructing, maintaining, and rationalizing complex organizations of documents. Similar findings are reported in a study comparing folders and tags as methods of organizing personal information (Civan, Jones, et al., 2008).

A final possible reason that filers access proportionally less of their data is that they simply have more stuff. There are finite constraints on how much data one can access. Filers have more data and, as a consequence, they are able to access less of it. This is consistent with the observation that the absolute amounts of data accessed by both

groups were very similar.

We also expected filers to be quicker to rationalize their data in preparing for the move, given the greater care they had initially taken to organize their data. But there were no differences in packing time for filers and pilers. This could be because filers' greater organization is offset by having more data through which to sift. And contrary to our predictions, pilers found it subjectively easier to rationalize archives in preparation for the move. Why was this? Even though filers discarded more reference information, they generally found it difficult to discard filed documents, partly because of the investment they had already made in managing that information. Filers therefore seemed less disposed to discard information they had invested effort in organizing. In contrast, unfiled information seemed easier to discard.

Finally, we looked at what determined strategy choice. Although job type influenced strategy somewhat (e.g., secretaries were more likely to be filers), in general strategy seemed to be more affected by dispositional factors.

<B> **Managing Digital Files and Folders**

We access our files and folders on a daily basis and their organization has clear importance for our everyday digital lives, yet there have been relatively few studies of how people organize these files and what affects this organization. One exception is Boardman and Sasse's (2004) study, which looked at the structure of people's personal data; they found that on average people had 57 folders with a depth of 3.3 folders. That study also documented different filing strategies, finding that 58 percent of people systematically filed information items when they created them, a further 35 percent left

many items unfiled (in a manner similar to paper piling), with a small proportion (6 percent) leaving most items unfiled. In some cases, people did not file actionable documents (i.e., those one which they were currently working), instead leaving them in obvious places such as the desktop where they would be reminded about them. Boardman and Sasse also looked at the types of folders that people created, identifying two main classes: project and role oriented. Finally, they looked longer term to see whether management strategies changed over time but found little evidence for this.

Two other studies looked at the structure of people's file systems. Gonçalves and Jorge (2003) studied the folder structure of 11 computer scientists using Windows (8), Linux (2) and Solaris OS (1). Their results show extremely deep, narrow hierarchies. The average directory depth was 8.45, with an average branching factor (an estimation of the mean number of subfolders per folder) of 1.84, indicating a deep and narrow hierarchy. In contrast, a larger scale study by Henderson and Srinivasan (2009) looked at the folder structure of 73 university employees using Windows OS. The structures they found were much shallower, being only 3.4 folders deep on average. Folders tended to be broader, with an average of 4.1 subfolders per folder for non-leaf folders. Both studies found relatively small numbers of files per folder: 13 for Gonçalves and Jorge (2003) and 11.1 for Henderson and Srinivasan (2009).

In another study probing why people generate specific folder structures, Jones (2005) interviewed people about the nature of their folder systems. Consistent with external cueing, many folders were seen as *plans*—structures that people used to organize their future work. Folders represented main tasks and subtasks of ongoing projects, serving to remind people about aspects of their work activity that needed to be executed.

People also used workarounds to make various types of information more salient, for example, labelling folders “aacurrent” instead of “current” to ensure that this information was more obvious when browsing an alphabetically ordered folder list.

Bergman and colleagues (2003, 2009) also document workarounds *within* folders, to make individual files and folders more salient while avoiding the need to delete information. They describe how people create subfolders for older, less relevant information and label these *archive* or *old* to reduce clutter and make relevant working items more visible in the main active folder.

Another important aspect of digital file organization is the *adaptive* nature of active folders. Bergman and colleagues (2008) showed that the most common strategy for accessing personal information is navigation through the folder system, with this type of access occurring many times per day. One implication of this continual re-access is that users are likely to discover suboptimal organization, leading them to modify their file and folder structures. Adaptive maintenance and modification will turn out to be important when we discuss archives that are much less frequently accessed, which often prove to be poorly structured. For example, people add an average of 5.9 new files to their work collection each day, creating a new file folder every three days. In contrast, with digital pictures, months may elapse between the creation of new folders, negatively effecting people’s ability to retrieve those pictures (Whittaker et al., 2010).

More recently, new tools have been developed to support different types of organization. One example is *tagging*. The Phlat system (Cutrell, Robbins, Dumais, & Sarin, 2006) allows users to apply multiple labels to a given information item, rather than storing it in a single folder location. Tagging has the advantage of providing richer

retrieval cues (because multiple labels are available as retrieval terms) as well as allowing users to filter sets of retrieved items in terms of their tagged properties (e.g., 'pictures' + 'personal' returns files with those tags). In contrast, current file and folder systems are more restricted in how data can be accessed and navigated. If a file is stored in the *work2008* folder, unless I can recall or navigate to that exact folder location, I will be unable to relocate the data. But in spite of these putative advantages, in a long-term field trial users made very little overall use of tagging, averaging only one query per week with the Phlat system. It seemed from user comments that the costs of creating tags may have been too high to generate sufficient tags to support flexible search and filtering. This led people to use the system more like a standard desktop search tool. Another study compared tagging and foldering, again failing to find clear benefits for tags (Civan et al., 2008). In the next section we discuss how social tagging may reduce some of the costs of creating personal tags.

<B> **Managing E-Mail**

<C> **Actionable Items**

Managing e-mail is complex and different from paper or standard digital files. A critical aspect of e-mail is that it contains many actionable messages. To be effective, people need to organize actionable information in such a way that they are reminded of what they need to do and when. This means that users have to organize action-oriented information so that they will encounter it when they need to do so. We first describe *how* users process actionable messages. We then turn to what they do with informational

messages, which are treated more like paper and standard digital files.

For actionable items, deferral is inevitable. Only a small proportion of actionable messages can be dealt with at once; most must wait to be processed. Dabbish and colleagues (2005) found that on average 37 percent of messages that require replies are deferred, which equates to about 4 deferred messages per day. Forgetting these deferred tasks can create major headaches both for the user and the organization.

Whittaker and Sidner (1996) found that the most prevalent strategy for reminding about actionable messages is to leave them in the inbox. Users know that they will return to the inbox to access incoming unprocessed messages and thus perhaps be reminded about their outstanding actionable messages. Dabbish and colleagues (2005) also report that actionable items are left in the inbox around 79 percent of the time. We called this strategy *no filing*.

Whittaker and Sidner also showed the importance of using the inbox to prompt visual reminding by observing the failure of other strategies: 25 percent of users had experimented with a strategy of filing actionable items in a to-do, folder. In fully 95 percent of these cases, this folder was abandoned because people had to remember explicitly to go to it, open it, and review its contents. This extra effort contrasts with being reminded about outstanding actions merely by seeing them in the inbox when reading new e-mail. Although other studies (Bellotti et al., 2003) suggest that some users change their work practices to exploit to-do folders, this demands extra cognitive steps. Paradoxically, these users have to remember actively to look for their reminders. In contrast, items in the inbox are encountered naturally as a side effect of accessing new messages.

Of course, there are also disadvantages to leaving actionable items in the inbox: These reminders may be difficult to spot if the user receives many new messages. Incoming messages visually displace older pending actionable items—requiring users continually to scroll through the inbox to ensure that these items are not out of sight and out of mind (Whittaker, 2005; Whittaker, Jones, & Terveen, 2002b; Whittaker & Sidner, 1996). Tang and colleagues (2008) looked at the proportion of the inbox that users had constantly visible, finding that on average only 25 percent of the messages were in view. The remaining 75 percent were not therefore serving as direct visual reminders for outstanding actions—compromising their ability to remind.

Other users try to keep the inbox clear by filing incoming actionable items in dedicated, task-related folders (Bellotti et al., 2005; Whittaker & Sidner, 1996). Whittaker and Sidner dubbed these people frequent filers and reported that 25 percent of users create such folders. There are obvious advantages to this strategy: Removing items from the inbox keeps it trim and also allows users to focus better on new and important information. However, these benefits may be outweighed by disadvantages: Users are required to create, maintain, and continually check these task folders. Failure to file appropriately can also have severe consequences, if one files important information and forgets about it.

A final strategy for actionable items is a hybrid of these. Whittaker and Sidner (1996) identified a final group accounting for 35 percent of their users who engaged in *spring cleaning*. These people would wait until huge amounts of information accumulated in their inboxes, making it hard to identify actionable items. They would then engage in extensive filing to rationalize the inbox. The process would be repeated

with the inbox gradually growing in size until another crisis brought on extensive filing once more.

What determines which strategy people choose when processing actionable e-mail messages? Whittaker and Sidner (1996) looked at the impact on strategy choice of organizational role and incoming volume of messages. Managers were more likely to receive greater volumes of e-mail but there was no evidence of a direct relationship between strategy and role. As with our paper study, it may be that dispositional factors are an important determinant of strategy choice. This is supported by other research that demonstrated relations between cognitive style and strategy (Gwizdka, 2004a, 2004b).

Other studies of e-mail have found some support for these management strategies (Bellotti et al., 2005; Dabbish et al., 2005; Fisher et al., 2006; Mackay, 1988; Whittaker, 2005; Whittaker et al., 2002a). However, later work indicates few instances of pure no filers: people with absolutely no folders who are totally reliant on their inboxes for task management. Bälter (2000) both extended the set of management strategies and also argued that people move sequentially from being an active filer to spring cleaner, and later no filer, as the volume of e-mail they receive increases. He argues that those receiving the highest volumes of e-mail are those with these least time to organize it.

<C> **Informational Messages**

We now look at how users organize *informational* messages. A substantial percentage of e-mail messages are informational as opposed to actionable (Dabbish et al., 2005; Whittaker & Sidner, 1996). Users also experience problems in processing informational e-mail messages. Observations of e-mail behavior show that users spend

huge amounts of time organizing e-mail messages: On average 10 percent of people's total time in e-mail is spent filing messages (Bellotti et al., 2005).

Again Whittaker and Sidner (1996) examined why users have problems with filing such information. Creating folders for informational messages is hard for several reasons. Generating and maintaining folder collections requires considerable effort. Filing is a cognitively difficult task (Lansdale, 1988). Just as with the keeping decision, successful filing is highly dependent on being able to envisage future retrieval requirements. It is hard to decide which existing folder is appropriate or, if a new folder is needed, how to give it a memorable name that will be appropriate for the retrieval context in which it will be needed.

Again, as we saw in the keeping section, another reason for not filing is that users want to use the *deferral strategy* and postpone judgments about the value of information. Users do not want to create archives containing information that later turns out to be useless or irrelevant. They are aware that creating overly complex archives may make it harder to access truly valuable information.

Furthermore, folders may not be useful after they are constructed. One may not be able to remember folder labels, especially when one has large numbers of older folders. Research combining multiple studies shows that people have an average of around 39 e-mail folders (Whittaker et al., 2007). When filing they therefore have to remember the definition of each and to be careful not to introduce duplication by creating new folders that are synonymous with pre-existing ones. Duplication of folders detracts from their utility at retrieval.

In addition, folders can be too small to be useful. A major aim of filing is to

coerce the huge number of undifferentiated informational inbox items into a relatively small set of folders, each containing multiple related messages. Filing is clearly not successful if the number of messages in a given folder is small. If a folder contains only one or two items, then creating it has not significantly reduced the complexity of the inbox nor gathered together significant amounts of related material.

Our data show that filing often fails: On average 35 percent of users' folders contain only one or two items. Later studies duplicated these observations, but finding a lower percentage (16 percent) of such failed folders (Fisher et al. 2006). These tiny failed folders do not significantly reduce the complexity of the inbox; moreover, they introduce the dual overheads of: (1) creating folders in the first place and (2) remembering multiple folder definitions every time there is a decision about filing a new inbox item. This cognitive overhead is illustrated in that the larger the number of folders a user has, the more likely that person is to generate failed folders containing only one or two items (Whittaker & Sidner, 1996). Of course, a small number of these failed folders may represent new activities that the user is planning (Bergman et al., 2003; Boardman & Sasse, 2004; Jones et al., 2005) but such planning cannot account for all of these tiny folders.

Folders can also fail because they are too big. When there are too many messages in a folder, it becomes unwieldy. And as the relationships among messages within the folder become more tenuous, the benefit of keeping them together is much reduced. With large heterogenous folders, it can be extremely difficult to collate related items or find a target item (Whittaker & Sidner, 1996).

Elsweiler, Baillie, and Ruthven (2008) looked at the impact of filing strategy on

users' memory for their e-mail messages. Frequent filers tended to remember less about their e-mail messages. This is consistent with our earlier observations about premature filing. Filing information too quickly can lead to the creation of archives containing spurious information; quick filing also means that users are not exposed to the information frequently in the inbox, making it hard to remember its properties or even its existence.

Thus, e-mail users experience cognitive difficulties in creating folders for informational messages. In addition, the payoffs for this effort may not be great: Folders can be too large, too small, or too numerous for people to remember individual folder definitions. In consequence, folders may be of restricted use either for retrieval or for collating related messages. As we have seen, some users finesse this problem: Instead of filing informational messages, they simply leave them all in their inbox. More recent work has tried to support this strategy by introducing new techniques such as thread-based viewers, which we describe in the technology trends section.

<B> **Managing Webpages**

Unlike e-mail, Web information is largely not actionable: Users may want to ensure that they remember to read a webpage, but in general there are no negative consequences for failing to do this.

One prevalent form of managing Web information is to bookmark encountered webpages. Numerous studies have looked into how people organize their bookmarks. Two early studies documented the number of bookmarks created as well as their underlying structure. For example, Abrams and colleagues (1998) found that 6 percent of

respondents had no bookmarks, 10 percent had 1 to 10, 24 percent had 11 to 25, 44 percent had 26 to 100, 14 percent had 101 to 300, and 2 percent had more than 300 bookmarks. And Boardman and Sasse (2004) found that people organized their bookmarks into an average of 17 folders. Another study (Bruce et al., 2004) observed further strategies people use for organizing useful Web information that they encounter. In addition to bookmarking, users might forward themselves a link in e-mail, print the page, copy the link into a document, generate a sticky note, or rely on memory.

More recent work with more modern Web browsers has revisited bookmarking. Aula and colleagues (2005) looked at people's bookmark collections and found that 92 percent have bookmarks, with an average of 220 links, although there is huge variance: 21 percent of people have fewer than 50 bookmarks and 6 percent have none. The largest collection contained 2,589 links with 425 folders. Most of Aula and colleagues' (2005) informants reported major problems in organizing and managing their collections. Consistent with other studies (Tauscher & Greenberg, 1997) users often bookmarked information that they never subsequently revisited. In contrast, other studies showed that users were unwilling to create new bookmarks, fearing that creating bookmarks for information of unclear utility would clutter their existing set of useful bookmarks—compromising the utility of useful items (Aula et al., 2005; Wen, 2003). Aula and colleagues also found that the key for success with complex bookmark collections is the extent to which users actively exploit and maintain their collection of links. A subgroup of heavy users of bookmarks had collections of over 500 links; these users tended (like e-mail spring cleaners) to clean up their collections from time to time, deleting unused or no longer functioning links. They also carefully organized bookmarks into hierarchical

levels (similar to a file system). For these users who invested organizational effort, bookmarks seemed to be an indispensable tool. Abrams and colleagues (1998) also looked at the types of strategies people used for organizing their bookmarks. They found four main types: About 50 percent of people were sporadic filers, a further 26 percent never organized bookmarks into files, around 23 percent created folders when they accessed a webpage, and around 7 percent created folders at the end of a session. Creating folders also seems to be a response to having too many bookmarks on a drop-down list, so that people with fewer than 35 bookmarks have no folders but, beyond this threshold, folders grow linearly with the number of bookmarks.

Some disadvantages of bookmarking relate to the costs of creating and maintaining collections, especially as information needs change. Recent social tagging systems, such as Deli.cio.us, Dogear, Onomi, and Citeulike, may finesse some of these problems. These social tagging systems allow users to create multiple labels for the same data, providing potentially richer retrieval cues (Cutrell, Robbins, et al., 2006; Lansdale, 1988). More importantly, they allow tags to be shared among users, reducing the cost of tag creation for each user. Of course, the approach raises important questions. Do different users agree on a common classification of information or do they generate inconsistent, orthogonal tag sets? Numerous studies have shown that, given sufficient numbers of users, tag sets tend to stabilize on common descriptions of Web resources so that people can exploit others' tags (Golder & Huberman, 2006; Millen, Yeng, Whittaker, & Feinberg, 2007). Furthermore, with suitable user interface design (e.g., text completion) problems such as inconsistent spellings can be finessed and promote greater awareness of others' tags (Millen et al., 2007). If enough people are prepared to tag,

social tagging seems a useful tool that removes some of the costs associated with standard, individual bookmarking methods.

<B> **Managing Photos**

Photos are very different from e-mail messages and webpages, tending to be *self-generated* (like many files), and are usually neither informational nor actionable. They are also perceived to be highly important and often irreplaceable (Petrelli, Whittaker, & Brockmeier, 2008; Whittaker et al., 2010). How, then, do people organize them? Recent studies show that people manage to organize photos using rather rudimentary structures (Kirk et al., 2006; Whittaker et al., 2010).

Whittaker and colleagues (2010) investigated how parents organized family photo archives. They found that these collections tended to have very little hierarchical structure and were organized more like piles than files. Participants typically relied on a single main picture storage location (such as the "My Pictures" folder). For participants with multiple computers or external hard drives there was usually a single main storage folder for each device. People usually stored their pictures in that location in a single-level, flat hierarchy with minimal subfolders. Furthermore, when a target folder was opened and scanned, the folder often contained heterogeneous data, comprising pictures that related to multiple events (possibly because they were uploaded at the same time and never subsequently reorganized).

How can we explain this lack of organization? Previous work has highlighted how participants are able to exploit their familiarity with *recently taken* pictures to scan, sort and organize materials for sharing with others quickly (Kirk et al., 2006). Possibly

because of these experiences with recent pictures, participants may have expected themselves to be very familiar with their *entire* picture collection, and as a result were not motivated to organize their collections carefully. In most cases, it seemed that people had not accessed the vast majority of their pictures since they were uploaded. We saw evidence of this during retrieval. Participants universally preferred to view pictures in the thumbnail view for easier scanning. Had the participants previously opened these folders, we would have expected to see thumbnails. Yet during retrieval, when participants first opened their folders, photos almost always appeared in the "list" view, suggesting folders had rarely been accessed. And because participants seldom accessed pictures, they did not discover how poorly organized these were. One reason for the lack of organization and unfamiliarity is that parents typically have very little spare time to organize their photos. One participant commented that his attitude to photos was "collect now – organize later – view in the future."

Another way to organize might be to *annotate* pictures. However, consistent with earlier studies (Frohlich, Kuchinsky, Pering, Don, & Ariss, 2002; Kirk et al., 2006; Rodden & Wood, 2003), we found very little evidence of annotation. One reason is that annotating is onerous. Another problem, also observed in earlier studies (Kirk et al., 2006; Rodden & Wood, 2003), is that users may not annotate because they are *unaware* that they are likely to forget key aspects of pictures. People can currently remember detailed information about recent pictures and this may mean they have little motivation to annotate pictures for the eventuality that they will forget.

<B> **Management Summary**

Management is a difficult activity, because it requires people to predict when or how information will be accessed. To create effective organization, users have to anticipate the context in which they will be accessing information. And for action-oriented items, they have to anticipate exactly when they will need those items.

Information properties have a major impact on management strategy: *Actionable* items often require deferral, so people need to be *reminded* about them. Various tracking strategies facilitate reminding, including leaving actionable information in one's workspace, as well as using dedicated task folders. There are trade-offs between these strategies: Keeping information in a workspace affords constant reminding but it reduces efficiency because that workspace can become cluttered with many unrelated actionable items. A specific problem with using the e-mail inbox for reminding is that as new items arrive they tend to displace older actionable items putting them out of sight and out of mind. The disadvantage of dedicated task folders is that these need to be constantly accessed and monitored.

For *informational* items, people use two main strategies, filing and piling. There are surprising advantages for a paper piling strategy. Pilers manage to build up smaller archives, with more frequent access to information in the archive. In addition, we found problems with filing, including premature filing of low value information, leading people to generate complex collections of information that are of little utility.

For *informational* items, users experience difficulty in categorizing information, failing to predict accurately the context in which they will want to retrieve that information. People create folders that are both too big—containing large collections of heterogeneous items—and too small—containing one or two items in a folder that is

seldom used. People can also create duplicate folders for the same content. All this makes filing error prone.

Both users' dispositions and the volume of information they receive may influence the type of organizational strategy they employ. Users who receive large volumes of incoming information are under pressure to keep their workspaces clear (otherwise they may overlook important deferred actionable items) but they are the people who are least likely to have the time to file and organize their information.

Certain types of information, such as webpages and photos, are infrequently re-accessed. Infrequent access may mean that people fail to realize what information they have available and how poorly organized it is. Tags do not seem to be useful in the context of personal files but they do seem to have benefits in a Web/intranet context, where people can reduce the cost of annotation by sharing others' labels.

<A> **Exploitation**

<B> **Overview, Problems, and Strategies**

In this section, we first contrast exploitation with classic information seeking and foraging behaviors, then go on to describe different strategies for exploitation as well as the costs and benefits of these strategies.

<C> **Exploitation Not Information Seeking**

Exploitation is different from information foraging and classic information seeking. In both foraging (Pirolli, 2007; Pirolli & Card, 1999) and classic information

seeking (Belkin, 1980; Marchionini, 1995; Wilson, 1999), the target information is seen as being *totally new*. Exploitation is different in several ways. First, retrieval structures are usually *self*- rather than *publicly* generated (Bergman et al., 2003; Lansdale, 1988). In other words, people are searching their own organization and not a public database. Second, the exploiter may *remember* significant details about the target information item and how it has been organized.

For example, Gonçalves and Jorge (2004) asked participants to tell stories about three personal documents on which they had recently worked. People could remember a great deal about these documents with the most salient characteristics being age, location and purpose of the document. Blanc-Brude and Scalpin (2007) also found that location, format, age, keywords, and associated events were frequently remembered. Because people remember this information, access is not purely reliant on *external publicly provided* metadata ("scent" in the terminology of information foraging). Instead, it is mediated by *cueing*: where cues can be mental (the internal cognitive information users remember about the target before they begin to access it) or external (triggers provided by well-chosen folder or file names as users carry out their search). Indeed, as we saw earlier, management activities have the predominant purpose of constructing personal organizations that promote future exploitation.

Exploitation therefore involves reconstruction of partially familiar personally organized information, rather than evaluation of unfamiliar, publically organized data. A further difference concerns success criteria: While seeking information, it is often enough to access information that satisfies certain general properties ("cheap flights to Spain"), where multiple documents may satisfy this search. In contrast when accessing personal

information, the user often has a specific document in mind—making the criterion for success much more stringent. Of course, such prior knowledge may make retrieval easier. During access, users may quickly recognize the target document, so they do not have to scrutinize it to determine its relevance as they would an unknown webpage. But in other ways, access to very specific information can be made harder when it is satisfied only if a specific item is found; and there may be strong feelings of frustration about failure to locate that item (Whittaker et al., 2010).

<C> **Exploitation Strategies**

Exploitation success depends on the match between cues/structures generated for future retrieval and the extent to which they match that future retrieval context. Note that even if people rely on search, they still have to *generate* the relevant search terms to guarantee success; this requires them to reconstruct important aspects of the target document (e.g,. title, keywords, date). If there is a good match between organizational cues and the retrieval context, retrieval will succeed. But to create effective retrieval cues, users need to anticipate successfully *when* and *how* they will consume information.
We access personal information in four main ways.

One very straightforward way to access information is to *navigate* for it. For information items such as files, we navigate within self-generated hierarchies of folders and subfolders. People usually traverse their organizational hierarchy manually. They visually and recursively scan within each folder (either actively by sorting the items by attribute or by using the system default) until they locate the folder that contains the target item.

Search is another way to access personal information. An important emerging technology for exploitation is desktop search, allowing users to locate information from within their own file systems, using keyword queries, in the same way they conduct Web searches. First the user generates a query by specifying some property of the target item, including at least one word related to the name of the information item, and/or the text that it contains (full text search), and/or any metadata attribute relating to that item (e.g., the date it was created). The desktop search engine then returns a set of results from which the user selects the relevant item. Search has elsewhere been characterized as a form of teleporting whereby users move directly to the target information, without the intermediate steps that characterize navigation (Teevan, Alvarado, Ackerman, & Karger, 2004).

A third access method, *orienteering* is a hybrid combining both navigation and search (Teevan et al., 2004). When orienteering, users may generate a search query to locate a particular resource page or folder and then manually navigate to the target; or they might begin by accessing a link and use information from that link to generate a new search query.

Finally, new technologies such as *tagging* allow users to apply multiple labels to an information item whether on the desktop (Cutrell, Dumais, et al., 2006) or on the Web (deli.cio.us, Flickr). This allows users more flexibility in how they categorize the item (more than one label can be applied). Multiple tags mean richer retrieval cues, because the same information can be accessed via several different tags.

These strategies apply to *personal* information. When people incorporate public information into their personal schemes (e.g., Web bookmarking or history lists) more

varied strategies are possible (Aula et al., 2005; Bruce et al., 2004; Jones, Bruce, & Dumais, 2003; Obendorf et al., 2007). For example, users can deliberately bookmark valued information or save it to disk and then navigate back to the data. Or they can apply less effortful strategies such as accessing information via the history list (a list of sites visited), or use the browser's back button to re-access recent information.

<C> **Costs and Benefits of Exploitation Strategies**

If the fit between the organization that users construct and the retrieval context is inexact, even careful management strategies may not guarantee successful retrieval. The wrong classification of information can hide it from the user, reducing the chance of quick retrieval (Kidd, 1994; Malone, 1983; Whittaker & Sidner, 1996). Putting information in a folder may decrease its ability to remind, which may be vital for actionable information. In addition, because categorization is itself cognitively challenging, users may create spurious folders that are seldom accessed, which may make classification of new information harder (Fisher et al., 2006; Whittaker & Sidner, 1996).

What, then, are the trade-offs between navigation and search for accessing personal information items? There are clear benefits to navigation. Accessing information using a personally constructed organizational hierarchy is predictable and includes a spatial component that users find valuable (Barreau & Nardi, 1995; Bergman et al., 2008; Jones & Dumais, 1988; Robertson et al., 1998). Access takes place in incremental stages, so that users obtain rapid feedback about the progress of their access efforts, being able to backtrack if they find they have followed the wrong branch of their file hierarchy. At the same time, there are disadvantages to navigation, compared with search. In complex

organizational structures, navigation can be inefficient; and taking a wrong step early in the access process may require extensive backtracking, depending on the precise nature of the organization scheme (Hearst, 1999). Furthermore, users have to remember at retrieval time, how information was classified, which can be difficult when there are multiple categorization possibilities (Lansdale, 1988; Russell & Lawrence, 1997).

There are also potential advantages of *search* when accessing personal information. Search does not depend on users remembering the exact storage location or precisely how they classified their information; instead, they can specify in the query any attribute they happen to remember (date, name, filetype) (Lansdale, 1988). Search may also be more efficient: Users can potentially retrieve information in one step, via a single query, instead of using multiple operations to navigate to the relevant part of their folder hierarchy. More radically, search also has the potential to finesse the management problem*,* as users do not have to apply organizational strategies that exhaustively anticipate their future retrieval requirements.

The same dichotomy between navigation and search does not apply to actionable items. Here very different strategies must be used. *Reminding* is key; information must be organized in such a way that users encounter it opportunistically. Neither search nor navigation through complex file organizations is appropriate support for actionable items; both require *deliberate* acts to seek out data, whereas the primary characteristic of actionable items is that these should trigger *automatic reminding.* This is clearly a very hard problem: Effective reminding means users do not just want to *re-encounter* actionable information, they want to see it *exactly when or where they need it*. Actionable information presented at the wrong time may be highly distracting; it turns out that very

different strategies are needed for actionable than informational items.

For public data (e.g., from the Web) that people want to incorporate into their personal organizational schemes, it is apparent that users may have less incentive to manage public data because it is less highly valued, being less personally relevant or unique (Boardman & Sasse, 2004; Whittaker & Hirschberg, 2001). There are also clear trade-offs between different exploitation strategies for public data (Bruce et al., 2004). Although browsers now offer support in the form of suggestions, regenerating prior searches still requires considerable effort in remembering search terms, especially because search is often iterative—involving multiple queries relating to a specific information need, some of which may result in dead ends (Morris, Ringel Morris, & Venolia, 2008). Retracing successful navigation is also hard. Users have to remember which links they traversed. Bookmarking requires people to remember which information they have bookmarked, as well as to maintain bookmark collections. And more passive strategies, (e.g., relying on the history list) means that users have to navigate through poorly structured traces of every piece of information they accessed rather than just information that they thought was valuable (Morris et al., 2008; Wen, 2003). In all cases, retrieval may be made more difficult by the changing nature of the Web, which may alter the content of previously accessed pages.

We now discuss different strategies that people choose for exploitation of different types of information: namely files, e-mail messages, photos, and Web information.

<B> **Accessing Files**

Desktop search has seen significant recent developments. One limit of older search engines, such as those provided as part of the Windows and Macintosh operating systems, is that they allow users to search only one data format at a time. Following the Stuff I've Seen (SIS) initiative (Dumais et al., 2003), newer search engines support multiple formats—files, e-mail messages, instant messages, and Web history can be accessed within the same search query. They therefore have the potential to address the project fragmentation problem—where information items related to the same project are automatically stored in different locations, often because they depend on different applications (Bergman et al., 2003; Dragunov, Dietterich, Johnsrude, McLaughlin, Li, & Herlocker, 2005). Modern search engines are also substantially faster than older ones, with more sophisticated interfaces to specify their search choices (Farina, 2005; Lowe, 2006). Search is now also *incremental*, returning results as soon as the user begins typing the query. This incrementality allows users to refine their query in light of the results returned and truncate the query after typing just a few characters if the target item is already in view.

In a recent study (Bergman et al., 2008), we investigated whether advanced desktop search was replacing navigation as the main method for file access. We used multiple methods (longitudinal evaluation, large-scale cross sectional surveys) and examined different search engines (Windows XP search, Google Desktop, Mac Spotlight, Mac Sherlock). Users reported how often they searched versus navigated to their files. We verified the accuracy of the self-report data by collecting logfiles that allowed us to correlate self-report data with actual behavior. Self-reports were very accurate and highly correlated with actual behavior, with statistical correlations being around 0.94.

We know that organization requires effort—having to create and maintain appropriate structures that anticipate retrieval, as well as having to remember those structures during exploitation. Given these new search engine capabilities, we expected users to shift from relying on navigation for file access and become increasingly reliant on desktop search. We expected that people having access to desktop search engines with advanced features would be more likely to access their files using search than those who were using older search engines without those features.

Contrary to our expectations, we found that navigation was still users' preferred method for accessing their files. First, regardless of search engine properties, there was a strong overall navigation preference: Users estimated that they used navigation for 56 to 69 percent of file retrieval events and searched for only 4 to 16 percent of events. The remaining accesses were when users relied on shortcuts or used recent files to access items on which they had been working. Further, the effect of improving the quality of the search engine on search usage was limited and inconsistent. Although Google Desktop (which was fast, incremental, and supported cross-format search) led to more usage than Windows XP search, there was no evidence that other, more advanced features induced greater usage. For example, both Mac search engines were used equally often, even though the later version, Spotlight, was faster and supported cross-format, incremental search. Similar results using very different qualitative methods have also shown that pure search is uncommon. Instead, users often combine search with navigation (Teevan et al., 2004).

How can we explain why retrieval strategy seemed to be largely independent of search engine quality? One reason is that search often seemed to be used as a last resort

when users could not remember a file's location. Bergman and colleagues (2008) asked users to characterize exactly when they used search as opposed to navigation and found that between 83 and 96 percent of the times when people searched, they did so because they were unable to remember the files' location. When they can remember, they rely on navigation.

It also seems that in the majority of cases users can remember where files are located. This is unsurprising if we think that for common tasks we are frequently accessing and modifying information related to specific, often recent, items (Dumais et al., 2003) and this reinforces our memory of those items and their locations. As we have seen, people are able to remember substantial amounts of information about recent files (Blanc-Brude & Scalpin, 2007; Gonçalves & Jorge, 2004). The conclusion that search is used only when people cannot remember the location of a file is supported by other studies. Jones and colleagues (2005) found that only 7 percent of users were happy with the idea that they could dispense with folders even when desktop search was available.

<B> **Accessing E-Mail**

Accessing information in e-mail is a critical problem, given the amount of time that people spend processing it and the fact that it is both a to-do list for actionable information as well as an archive for more informational data (Duchenaut & Bellotti, 2001; Whittaker, 2005; Whittaker & Sidner, 1996).

A critical aspect of e-mail management is to ensure that actionable items are dealt with to meet specific commitments. The previous section noted that the most common reminding strategy is to leave such items in the e-mail inbox, hoping that these will be re-

encountered on returning to the inbox to process new incoming information (Bälter, 2000; Bellotti et al., 2005; Dabbish et al., 2005; Mackay, 1988; Whittaker, 2005; Whittaker & Sidner, 1996; Whittaker et al., 2007). Variants of the inbox as to-do list strategy include altering the status of actionable items that have been read and resetting the status of such messages so that they appear to be unread and hence bold in a standard browser (Whittaker, 2005).

In spite of the central role of e-mail in everyday work, we know relatively little about how people actually retrieve information from e-mail. One exception is a study by Elsweiler and colleagues (2008) who looked at people's ability to remember e-mail messages. Participants were usually able to remember whether a message was in their collection. Also memory for specific information about each message was generally good, with users often remembering multiple attributes. People remembered content, purpose, or task-related information best, correctly recalling over 80 percent of this type of information—even when items were months old. They were less good at remembering sender information; memory for this type of information tended to decay rather quickly. Memory for temporal information was worst of all, dropping to around 50 percent correct over several months. In all cases, memory was affected by both the age and size of the e-mail archive, with users remembering less when they had bigger archives or when they were required to remember older items.

Dumais and coelleagues (2003) also examined e-mail access in Stuff I've Seen. SIS is a cross-format search engine allowing users to access files, e-mail messages, webpages by issuing a query in a single interface. It also supports sorting of results via attributes such as date or author. The majority of searches (74 percent) was focused on e-

mail as opposed to files. This may be because, as we saw earlier (Bergman et al., 2008), if people want to access files, they do so using navigation rather than search. When searching for e-mail messages, there was a very strong focus on recent items, with 21 percent of searched-for items being from the last week and almost 50 percent from the last month. Many of these searches (25 percent) included the name of the e-mail sender in the query, suggesting (contrary to Elsweiler et al., 2008) that sender name is a useful retrieval cue for e-mail messages. Elsewhere we exploit the salience of sender name in the ContactMap system, which provides a specific informational view of e-mail data, centered around network models of sender data (Whittaker, Jones, Nardi, Creech, Terveen, Isaacs, et al., 2004). How can we explain the prevalence of name-based search observed by Dumais and colleagues, when compared with Elsweiler and colleagues' (2008) results? Part of the difference may be due to the Dumais group's observations of naturalistic behaviors, which tended to be focused around retrieving recent e-mail messages. In contrast, the Elsweiler team looked at longer-term access, for more structured, lab-based tasks. In addition, Dumais and colleagues did not look at the success of searches; it may be that although sender information was used frequently in searches, these sender searches were often unsuccessful.

<B> **Accessing Photos**

We have already described how people organize their digital pictures and the rudimentary management strategies that they employ. As with e-mail research, there has been more focus on photo management and rather less examining exploitation. Digital photos are a highly valued resource (Petrelli et al., 2008; Whittaker et al., 2010), so we

should expect people to create effective ways to access them. Indeed, work on accessing recently taken photos shows that people are good at retrieving these (Frohlich et al., 2002). When Kirk and colleagues (2007) asked people to sort recent pictures in preparation for sharing them with friends or family, they found that participants were effective in finding and organizing pictures taken within the last year.

These findings contrast with our own work on parents' ability to retrieve slightly older family pictures (taken more than a year ago). Although pictures were judged as being highly valued, participants were often unsuccessful in accessing such older pictures.

We asked participants to name significant family events from more than a year earlier that they had photographed digitally. In a subsequent retrieval task, participants were asked to show the interviewer digital pictures from 3 to 5 of these salient past events concerning their children. To prevent participants from choosing events that they could retrieve easily, they were not told about the retrieval task during the initial interview. The interviewer asked participants to sit at their computers and show him pictures relating to these key events.

In contrast to their expectations, our participants were successful in retrieving pictures in only slightly more than half of the retrieval tasks (61 percent). In the remainder (39 percent), participants simply could not find pictures of significant family events. Of the 28 unsuccessful retrieval tasks, 21 (75 percent) were pictures that the participants believed to be stored on their computer (or on CDs) but which they subsequently could not find. The remaining seven were pictures participants initially thought were stored digitally but during the retrieval process changed their minds into

thinking they were taken with an analog camera.

Based on participants' comments and behavior during and after search, we identified several potential reasons for their unexpectedly poor retrieval performance: too many pictures, distributed storage, unsystematic organization, false familiarity, and lack of maintenance. In our discussion of management we have already talked about the absence of systematic organization and the tendency to collect too many pictures; we now explore the implications of these for retrieval.

The most frequent explanation participants gave for their retrieval difficulties was that they had very large numbers of pictures to search. Consistent with previous work (Frohlich et al., 2002; Kirk et al., 2006; Rodden & Wood, 2003), participants felt that they were taking many more digital pictures than they had with analog equipment. All participants pointed to the low cost of capturing large numbers of digital pictures. However, during retrieval they realized that having too many pictures has its price when this mass of pictures competed for their attention, making it hard to locate specific ones. Average archive size was 4,475 pictures but with huge amounts of variation (SD 3,039). This is a striking finding because, consistent with other research (Kirk et al., 2006), participants all made definite efforts to reduce the overall number of pictures. For example, they deleted around 17 percent of poorly focused or unwanted pictures, both when pictures were first taken, as well as at upload.

Some participants attempted to account for their poor retrieval by arguing that they had not given folders meaningful names. However, 67 percent of participants made efforts to apply meaningful labels rather than relying on software defaults. But this did not seem to guarantee they could find their pictures, possibly because, as we saw in the

management section, naming schemes were inconsistent. People who used meaningful labels were neither more successful nor faster at retrieving pictures. Participants' comments and behaviors also suggested that the meaning of such names was sometimes forgotten over time. Finally, participants commented on difficulties in remembering changes over the years in organizational schemes they had imposed or software they had used.

The lack of organization in people's collections meant that they were over-reliant on trial and error strategies for accessing their photos. Consistent with studies of autobiographical memory (Brewer, 1988; Wagenaar, 1986), some of our 18 participants tried to use knowledge of related events to remember the *approximate date* when the target event occurred and then navigate using date information to the folders they thought might contain these pictures. Specific folders were chosen because their names (if there was a meaningful name) was thought to relate to the target or because a folder date was close to the guessed date.

Others tried to remember the exact date when the event had occurred and to find folders from that date. This worked when folders had been labelled with correct dates, although in many cases folder labels were purely textual. We have already noted problems with this strategy. First, participants may be unable to remember the date of the target event accurately. Second, the date label itself may be inaccurate, either because of problems with camera settings or the folder date representing the upload date as opposed to when the picture was actually taken.

Overall, the retrieval strategy used most often seemed to resemble trial and error: Users would cycle through their entire photo collection, accessing folders to see whether

they contained promising pictures and moving on to other folders if they did not.

<B> **Accessing Web Documents**

The problems of accessing webpages have been much studied. Most people's intuitions about Web accesses are that these follow the pattern of foraging: That people predominantly seek out *new* information from the Web, which they then consume for the first time. These intuitions also lead people to think that we typically rely on *search* to access Web information.

One possible reason for this belief in the dominance of search is that, historically, Web tools moved from relying on navigation via human-generated categories to being search-based. Early Web tools such as Yahoo! provided human-generated taxonomies of the then relatively small collection of Web documents, supporting access by allowing users to navigate through these hierarchies. One limitation of these manual taxonomic techniques is that they are completely impractical for the billions of documents that are now estimated to be on the Web. Self-report studies also suggest that usage of Web navigation is now much less frequent, with people reporting a far greater reliance on search for foraging (Kobayashi & Takeda, 2000).

In reality, however, it turns out that search is less frequent than we might expect. Instead of foraging for new information, users tend to re-access previously visited data using a variety of simple browser techniques including following links, retyping the URL, or exploiting the back button (Aula et al., 2005; Bruce et al., 2004; Obendorf et al., 2007).

Many studies have attempted to document the extent to which Web accesses

involve information seeking versus refinding by analyzing logfiles and history lists. Early work looking at students' browsing behaviors showed that a characteristic Web access pattern involved hub-and-spoke accesses, in which users would find a useful authoritative resource—a hub. They would then fan out to the various links from this page (spokes), usually traversing no more than two links before re-accessing the hub using the back button (Catledge & Pitkow, 1995). Tauscher and Greenberg (1997) instrumented browsers and looked at the rate at which people returned to previously visited sites. They documented a recurrence rate of 58 percent, finding also that the majority of overall accesses targeted a small set of websites that the user frequently re-accessed. Revisits are prevalent, as indicated by the use of the back button, which accounts for around 30 percent of Web actions. In addition, Tauscher and Greenberg found that people were much more likely to re-access sites that they had visited recently. Cockburn and Greenberg (2000) carried out a similar study, finding that a much higher frequency of accesses (81 percent) were revisits.

Another study conducted by Wen (2003) was unusual in looking at the *success* of refinding. He asked users to conduct typical Web access sessions and then requested them to retrieve information that they had found useful in that search session. Users were able to re-access successfully only 20 percent of the sites they had visited. They often failed to bookmark useful information, believing that doing so would create clutter and compromise their existing bookmark collections. Finally, and consistent with other results (Teevan et al., 2004), Wen found that the general strategy for re-access was to try to retrace prior actions, rather than attempting to search or type in prior URLs.

Aula and colleagues (2005) looked at users' self-reported strategies for Web

search and re-access. They found that having multiple windows or tabs open was very common because re-access was prevalent. In addition, the most commonly reported ways to re-access information were to: re-access links, search for it again, directly type the URL, or save pages as local files. This confirms the results of an observational study by Bruce and colleagues (2004) that documented that the most prevalent strategy for refinding was to type in the URL. Other access strategies were much less prevalent, for example, e-mailing links to oneself, adding URLs to a website, or writing down queries. Finally, there is very little use of history lists for re-access. Aula and colleagues found various problems with history lists: Not only are page titles often misleading, the list shows important and unimportant results intermingled—making it hard for users to focus on valued information. Both Aula and colleagues (2005) and Wen (2003) also noted user problems with re-access: in particular, using search to exploit information is difficult because it is an iterative process often involving multiple queries. Users may try multiple routes to finding information, exploring sites that later turn out to be dead ends. In trying to recover from these dead ends, users often could not regenerate previous accesses that had been more successful. Users also could not recall the exact method that they had used for access; as a result they had problems in reconstructing search queries for information for which they had originally browsed.

In perhaps the best controlled study of revisiting, Obendorf and colleagues (2007) preprocessed sets of URLs for 25 users and found that revisiting rates in prior studies might have been artificially inflated by sites that automatically refreshed without user intervention. When they controlled for such automatic refreshes, revisitation levels were around 41 percent. They also documented a variety of general strategies used to access

pages. The most common were: using a hyperlink (44 percent of accesses), using forms—including the use of search engines (15 percent), back button (14 percent), opening a new tab/window (11 percent), and typing in the URL directly (9 percent).

Turning specifically to revisits (as opposed to all searches), Obendorf and colleagues (2007) again found that the most common strategy for refinding information was to follow links (50 percent), with the back button being the next most common strategy (31 percent). The remaining direct access strategies (using bookmarks, homepage links, history, direct entry of URL) accounted for the final 13 percent of accesses. As in previous studies, re-accesses tended to be for recently visited sites: 73 percent of revisits occur within an hour of the first visit, which makes the reported use of the back button appear rather low. One possible reason for the relatively low numbers of back accesses may be that the tabbing facilities provided by new browsers mean that users are not as reliant on hub-and-spoke type re-accesses. They can, therefore, keep the context of their hub page while using tabs to manage follow-up spoke pages.

Finally, Obendorf and colleagues (2007) looked at how access strategies varied as a function of the length of time since the original page access. Again, there were huge recency effects, 50 percent of revisits occurred within three minutes and the dominant strategy here was to use the back button, presumably because the target information was readily available in the browser cache. For revisits occurring within the hour, the back button and links were the most common ways to refind data. Between an hour and a day, back button usage decreased hugely, with users becoming more reliant on links and direct access (typing in the URL). Between a day and a week, links and typing URLs were the most common strategies; and at intervals of greater than a week, use of links dominated.

This greater reliance on links may reflect an orienteering strategy (Teevan et al., 2004), in which users generate plausible sets of links and then choose among these for the final stage of access. In any case, the results clearly show that access strategies are quite varied and are heavily dependent on the time interval between initial access and re-access. Part of the reason for this is technical: For very short term re-accesses, information is directly available in the cache, whereas at longer intervals this is unlikely to be true. In addition, cognitive factors are at work here. At medium and longer re-access intervals, users may have generated several windows or tabs so they are unable to remember which of these they first used to access the data.

Finally, the majority of revisits (73 percent) occur within an hour, 12 percent between an hour and a day, 9 percent between a day and a week, and 8 percent at longer intervals. As we have seen, the time between accesses is a critical factor influencing retrieval and, because the majority of revisits is really short term, certain strategies (such as using the back button or link-based access) are prevalent overall.

To summarize, then, Web retrieval often involves re-accessing previously visited pages. Use of links, tabs, and the back button is prevalent for more recently accessed pages. Search tends not to occur very often. Users also tend to access a small number of sites and other research shows that familiarity also influences retrieval strategy (Capra & Perez-Quinones, 2005).

<B> **Exploitation Summary**

During exploitation, people's preference is for manual methods (folder navigation/following links), whether this is for regular files or Web data. Search is a less

preferred option, even for Web documents.

Search is not successful with personal photos (content-based techniques are weak and there is very little metadata); and people, therefore, have to rely on browsing, which turns out to be ineffective for older data in many cases.

E-mail messages are different from files: Search can be useful for *informational* items because people are able to remember certain information about messages (names/content), at least in the short term. However, reminding is needed for *actionable* items and search cannot be used because it is a deliberate act that implies the user has already remembered. Users therefore have to rely on scanning their inboxes, which is often inefficient because of the amount of heterogeneous information they currently contain.

In spite of people's intuitions, search is not the prevalent way to access Web data. Re-accesses are very common, with people using the back button or hyperlinks as their main re-access methods. Re-accesses are usually for recently accessed information and the re-access strategy depends on how recently the target item was last accessed.

Mismatches sometimes occur between retrieval structures and their exploitation. For photos, there seems to be a failure to create retrieval-appropriate structures, which occurs in part because these are not frequently accessed; as a result, retrieval is often unsuccessful for older materials. For e-mail messages, people spend large amounts of time creating folder structures that may not always be exploited. For Web documents, people often create structures (such as bookmark collections) that are not used because there are less costly ways to access information. They also fail to create structures that are useful.

Retrieval has clear regularities—there is a strong bias toward access of recent items, as well as a bias toward accessing a small number of items very frequently.

<A> **Future Research**

What, then, are pressing future issues for research into information curation? In particular, because technology is so important in this area, what impact will emerging technologies have on keeping, management, and exploitation?

<B> **Technology Trends**

<C> **Keeping**

Storage is now so cheap that we no longer need to delete items because they are consuming valuable space. One general shift will, therefore, be away from models where users delete information, either when it is first encountered or during later cleanups. Instead people will tend toward keeping everything (Jones, 2004; Marshall, 2008a, 2008b), but with interfaces that provide views onto what is important and valuable in the data.

There are clear advantages to this keep everything approach. We know that users find deletion cognitively and emotionally difficult; and they are also concerned that they will end up deleting valuable information (Bergman et al., 2009). Keeping everything means that these difficult decisions can be at least partially avoided, although the consequence is that we need new approaches to management and exploitation if users are not to be overwhelmed by kept data. In this spirit, we have begun to build user interfaces

that keep more (assuaging worries about deleting something valuable), but that privilege information that is valuable or important. For example, motivated by a study of users' current workarounds with files and folders, we built GrayArea (Bergman et al., 2009), which implements a two-tier view of each folder, with the main view showing critical documents. The secondary area (GrayArea) is for less important files, which are made less visually salient, but still potentially available. A user evaluation showed the utility of this interface compared with the standard Windows Explorer method of managing files. Of course, one problem with this approach is that it requires manual organization to generate two-tier views; we are exploring (semi-) automatic methods for learning distinctions between these two types of information in an attempt to reduce the burden of manual organization.

Other technical possibilities involve the direct application of machine learning to address the keeping decision. Indices and profiles could be built based on the structure and content of people's current e-mail, files, and Web documents. These could also include information about which items are accessed and changed most frequently. The data could be used to generate an interest profile for the user, which could then be applied to incoming e-mail messages or recently accessed webpages. If, for example, an incoming e-mail message or viewed webpage closely matches information that is already in the user's file system, this item would be a clear candidate for keeping. In contrast, an e-mail message bearing no relation to the user's interests is a good candidate for deletion. One problem with this approach, however, is that it might be very effective at recognizing positive candidates for keeping but rather less good for deciding what should be rejected. Automatically deleting information that is unrelated to the user's current

profile introduces various problems. Just because incoming information is unrelated to the user's current activity does not mean that it is irrelevant. Unrelated messages, files, or documents might just represent an exciting new opportunity, an emerging new area, or a potentially important new contact; they should not, therefore, be deleted.

<C> **Management**

Programs built to support management have a long history (see Whittaker et al., 2007, for a review), in particular in e-mail, where many systems try to file or filter incoming e-mail messages automatically or semi-automatically. This approach has various problems, however.

One critical problem is that users fundamentally do not trust machine-learning programs (Pazzani, 2000). People are concerned that important incoming messages might be misfiled. It is clear that, in spite of large improvements in machine learning helped by the existence of new corpora, programs are still errorful (Whittaker, Hirschberg, Amento, Stark, Bacchiani, Isenhour, et al., 2002; Whittaker et al., 2007). And although programs promise to classify documents into folders correctly and with relatively low error rates, we still lack vital empirical data about what error rates are acceptable to users. Until we know clearly whether users will at best tolerate 5 percent of misfiling, we do not know what quality our machine-learning algorithms need to be.

One response to the errors problem is to use semi-automatic methods. Here the system suggests to the user where a document might be filed and the user confirms or corrects this. This approach is well liked by machine learning advocates because it provides a way for the user to generate structured feedback on the algorithm (Whittaker

et al., 2004; Whittaker et al., 2007). But there is a downside: Unless the interface is well designed, so that suggestions and user feedback are handled in a lightweight manner, the effort of correcting system suggestions may be greater than manual filing. Feedback and suggestions need to be extremely subtle with good defaults, otherwise the purported solution (automatic filing) may require more effort than users' current manual filing practices.

Another, perhaps more promising approach might be to use public resources to organize personal data. For example, systems such as Phlat (Cutrell, Robbins, et al., 2006) and Dogear (Millen et al., 2007) use social tags to organize personal resources. For example, a document in my filing system may inherit tags that others have applied to that document in a public archive. This approach has the benefits that user-generated tags are often more appropriate than machine-generated ones; it also reduces the management costs to the individual user, who has access to rich tags without having to generate them. However, there are various unanswered questions here, such as how to weight the importance of personally generated versus social tags. In addition, as we have seen, many of the user's most important documents are unique, making it unlikely that public tags are available to describe them.

Yet another approach to automatic management is to analyze user activity to determine the importance of, and relatedness among, documents. A common intuition is that documents we access frequently are more likely to be important, as are recently accessed documents. The "my recent documents" shortcut in MS Windows capitalizes on the latter intuition, and more principled algorithms have also been built to capture more systematic aspects of recency (Tang, Lin, Pierce, Whittaker, & Drews, 2007). Other

systems have used social information to profile documents, so that resources that are frequently accessed by others are visually privileged over those that are less frequently accessed (Kalnikaité et al., 2008).

One specific area where machine learning might be extremely beneficial is for actionable items, which are often a user's greatest concern when processing e-mail messages. Work on analyzing e-mail content has been relatively successful in predicting whether a given message requires a response (Cohen, 1996). Annotating e-mail messages with this information and presenting it in the interface might be very useful in helping people keep track of to-dos. Another approach to this problem is thread detection and visualization, which are now parts of newer e-mail clients (e.g., Gmail) and research prototypes (Bellotti et al., 2003; Tang et al., 2008; Venolia and Neustaedter, 2003; Wattenberg, Rohall, Gruen, & Kerr, 2005). These thread viewers attempt to reduce inbox 'clutter' by clustering related messages. This has the benefit of collating related information as well as reducing visual distraction in the inbox. Although there have been two small-scale evaluations of this technique (Bellotti et al., 2003; Tang et al., 2008), as yet we know little about how effective these techniques might be; one study (Tang et al., 2008), however, suggests that threading may interfere with established foldering practices.

Another specific area where we can expect developments in curation is with photos, where we have seen that users have major problems with management and exploitation (Whittaker et al., 2010). Standard metadata such as time and location might be supplemented with global positioning system (GPS) and compass data about where a camera is pointing (allowing inferences about what the shot might contain as well as

content-based tagging). GPS data might also indicate where a photo was taken (Kalnikaitė et al., 2010). And specific content-based techniques such as face recognition might allow familiar people to be tagged in pictures, a tool already available in Picasa and on the Macintosh. However, the promise of face recognition needs to be evaluated in the light of practical concerns. Name tags may be most important for infrequently encountered people whose identity the user is likely to forget; but will users be prepared to tag large numbers of people and will these programs work accurately for small numbers of relative strangers? And what about the success of these programs for people whose images change rapidly, such as infants and young children?

Machine learning has also been applied to task fragmentation. TaskTracer (Dragunov et al., 2005) is a system that analyzes user behaviors in an attempt to organize them according to activities. One major problem for users is fragmentation, whereby resources relating to a common project are placed in separate locations by applications. Thus the e-mail messages, spreadsheet, presentation, and documents for a project may all be in different folders, making it hard for users to collate and organize task-related materials (Bergman et al., 2003; Boardman & Sasse, 2004). TaskTracer addresses this by analyzing temporal access patterns: If a webpage, document, e-mail, and spreadsheet are repeatedly open at the same time, the system infers that they belong to the same task and constructs a virtual folder for that task. The user can choose to view resources in the virtual folder or in the regular file system, but the benefits of the virtual folder are that related materials are clustered together. Of course TaskTracer suffers from the same problems as many machine-learning programs in being imperfect but, because it is an alternative to manual files, users can employ it if and when it offers benefits.

<C> **Exploitation**

Technology might also be beneficial for various aspects of exploitation. One obvious area is desktop search. Although we have seen that desktop search is currently an infrequent way to access personal data, it is nevertheless potentially useful as a last resort (Bergman et al., 2008). One current problem is that desktop search typically generates too many irrelevant results. Search might be improved by including either social information (e.g., Millen et al., 2007) or more specific data about frequency and recency of document access.

Automatically captured data could also provide different ways to view and hence access our personal information. One approach might be to project different views onto the user's data, employing readily available metadata (time-based, social, location). These views are not meant to replace existing folders but to provide alternative ways to access their contents. For example, we have seen that usage information might be automatically time aligned, so that all resources accessed around the same time can be accessed together (Dragunov et al., 2005). Radical alternatives such as Lifestreams (Fertig, Freeman, & Gelernter, 1996a, 1996b) promise to replace our current semantic file systems with operating systems that are purely time based. Other radical approaches suggest that we might want to view all our information around social relations or social networks (Nardi, Whittaker, Isaacs, Creech, Johnson, & Hainsworth, 2002; Whittaker et al., 2004); these systems have also proved useful as alternative e-mail clients. Yet, other hybrid approaches combine search with key temporal events extracted from calendars or the public domain to allow people to access documents using these events as landmarks

(Ringel et al., 2003). For example, a user might be able to look at the personal information that was accessed shortly before a business trip to Boston or just after Thanksgiving, where the events are extracted from a personal calendar (the Boston trip) or public resource (Thanksgiving).

Such views could potentially be extended to other types of metadata. With the development of cheap sensors, it is now possible to record all sorts of information about what the user is doing at any time. Thus, it might be possible to provide information about where the user was when he or she worked on a document; and photos or other recordings might be available about other activities that the user was engaged in when that document was produced (Kalnikaitė & Whittaker, 2008b, 2010). For example, as a user might recall working on a presentation for a business trip to London; and a locational view might allow access to relevant documents by using this cue. Of course there are design challenges here: A huge amount of metadata available about users' activities already exists and interfaces will have to be carefully designed to ensure that the user is not overwhelmed by this richness.

<B> **Empirical and Methodological Issues**

One striking observation about information curation is that we know very little about it, in spite of its prevalence in everyday computer use (Whittaker, Terveen, & Nardi, 2000). Further, most previous research has focused on one aspect of the problem, namely management. We know much less about keeping and exploitation processes. This is somewhat ironic given the vast amount of research effort dedicated to systems and tools for accessing public corpora. More critically we do not know much about the

relationships among different aspects of information curation or, perhaps most importantly, how management strategies influence exploitation success. What, for example, is the relationship between a person's folder structure and his or her ability to retrieve and access files? Much more research is needed in this area. We also need to know more about when and why people keep or delete different types of information, exactly how they manage and reorganize, as well as the different methods that they use to access information.

Several practical reasons help explain why we know so little. First, it is extremely hard to gather data in this area. To understand information curation better, we need to collect data about people's personal information habits. This is potentially intrusive: It might require logging software to be installed on study participants' machines, or manual access to their personal data. And there are also problems with more system-oriented approaches: If we want to study the efficacy of new curation systems, these need to be both robust and fully featured. New curation software must be reliable because people use it on a regular basis for everyday work. If we want users to provide feedback about a new file system, e-mail client, or Web bookmarking system, that system had better be very effective or users will quickly switch back to their regular software. In the same way, the new system had better offer a comparable set of features to users' regular software, otherwise participants will quickly revert to that software to do their everyday work (Bellotti et al., 2005; Whittaker et al., 2004).

Further, methods for evaluating curation systems are complex and standard techniques cannot always be used (Kelly, 2006; Kelly & Teevan, 2007). For example, in evaluating information retrieval systems it is customary to use standard corpora and

measures such as precision and recall, where documents have been manually tagged for relevance. With curation systems, however, we need to evaluate systems against participants' own information because the use of public data would be meaningless. Further, users will generate their own access tasks exploiting their own management structures, so that methods relying on relevance metrics generated against standard corpora cannot be applied. In part this may explain why promising results obtained by the machine learning community using standard public corpora have not yet transferred well to practical curation systems. For example, new algorithms are able to categorize e-mail data in standard corpora with error rates around 10 percent. Yet we do not know: (a) what error rates users will tolerate for this type of task when carrying out everyday work; or (b) whether similar performance can be obtained with the user's own data. In our own work, we found that users were rather intolerant of automatic methods of clustering e-mail contacts, instead preferring semi-automated methods (Whittaker et al., 2004). More studies need to be carried out and better evaluation methods developed for information curation. Elsewhere we have advocated that the community develop a set of reference tasks for personal information management, which would allow comparative analysis of different algorithms across a common set of user tasks (Whittaker et al., 2000).

<A> **Summary**

This review has argued that prevailing views of information behaviors are misleading. Instead of being consumers of new public information, people's informational behaviors are closer to curation, in which they keep and manage personal information for future access. We have outlined a three-stage model of the curation

process, reviewing the central problems of keeping, management, and exploitation, and presented relevant data for each stage of the process, concluding with an overview of outstanding technical and empirical questions. In general, users tend to overkeep information, with the exception of contacts and Web pages. With respect to organizing information, we found surprising benefits for piles as opposed to files, although organizing action-oriented information remains a major challenge. Exploitation remains reliant on manual methods such as navigation, in spite of the emergence of desktop search. There are also mismatches between people's organizational structures and their actual retrieval requirements, for example, for e-mail, Web documents, and photos. Several new technologies have the potential to address important curation problems but implementing these in ways that users will find acceptable remains a challenge. Finally, research in this area remains in its infancy and new data and methods are still sorely needed.

<A> **Endnote**

1. Although Dabbish et al. (2005) suggest higher keeping rates for e-mail.

<A> **References**

Abrams, D., Baecker, R., & Chignell, M. (1998). Information archiving with bookmarks: Personal Web space construction and organization. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 41–48.

Ackerman, M. S. (1998). Augmenting organizational memory: A field study of Answer Garden. *ACM Transactions on Information Systems*, *16*(3), 203–224.

Ackerman, M. S., & Halverson, C. A. (2004). Organizational memory as objects, processes, and trajectories: An examination of organizational memory in use. *Journal of Computer Supported Cooperative Work*, *13*(2), 155–190.

Aula, A., Jhaveri, N., & Käki, M. (2005). Information search and re-access strategies of experienced Web users. *Proceedings of the International World Wide Web Conference*, 583–592.

Baddeley, A. D. (1997). *Human memory: Theory and practice*. Hove, UK: Psychology Press.

Bälter, O. (2000). Keystroke level analysis of email message organization. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 105–112.

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

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, *5*, 133–143.

Bell, G., & Gemmell, J. (2009). *Total recall: How the e-memory revolution will change everything*. New York: Dutton.

Bellotti, V., Ducheneaut, N., Howard, M., & Smith, I. (2003). Taking email to task: The design and evaluation of a task management centered email tool. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 345–352.

Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., & Grinter, R. (2005). Quality vs. quantity: Email-centric task-management and its relationship with overload.

Human-Computer Interaction, *20*(1–2), 89–138.

Bentley, F., Metcalf, C., & Harboe, G. (2006). Personal vs. commercial content: The similarities between consumer use of photos and music. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 667–676.

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. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 23–30.

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., Gradovitch, N., & Whittaker, S. (2008). Advanced search engines and navigation preference in personal information management. *ACM Transactions on Information Systems*, *26*(4), 1–24.

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. *Proceedings of the ACM International Conference on Human Factors in Computing Systems*, 269–278.

Blanc-Brude, T., & Scapin, D. L. (2007). What do people recall about their documents? Implications for desktop search tools. *Proceedings of the International Conference on Intelligent User Interfaces*, 102–111.

Boardman, R., & Sasse, M. A. (2004). "Stuff Goes into the Computer and Doesn't Come Out": A cross-tool study of personal information management. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 583–590.

Brewer, W. (1988). Memory for randomly sampled autobiographical events. In U. Neisser & E. Winograd (Eds.), *Remembering reconsidered* (pp. 21–90). New York: Cambridge University Press.

Bruce, H., Jones, W., & Dumais, S. (2004). Information behavior that keeps found things found. *Information Research*, *10*(1). Retrieved April 15, 2010, from informationr.net/ir/10-1/paper207.html

Capra, R., & Pérez-Quiñones, M. A. (2005). Using Web search engines to find and refind information. *IEEE Computer*, *38*(10), 36–42.

Catledge, L., & Pitkow, J. (1995). Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, *27*(6), 1065–1073.

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. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology* (CD-ROM).

Cockburn, A., & Greenberg, S. (2000). Issues of page representation and organisation in Web browser-revisitation tools. *Australian Journal of Information Systems*, *7*(2), 120–127.

Cohen, W. (1996). Learning rules that classify email. *AAAI Symposium on Machine Learning in Information Access*, 18–25.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Cutrell, E., Dumais, S., & 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. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 261–270.

Dabbish, L. A., Kraut, R. E., Fussell, S., & Kiesler, S. (2005). Understanding email use: Predicting action on a message. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 691–700.

Dragunov, A. N., Dietterich, T. G., Johnsrude, K., McLaughlin, M., Li, L., & Herlocker, J. L. (2005). TaskTracer: A desktop environment to support multi-tasking knowledge workers. *International Conference on Intelligent User Interfaces*, 75–82.

Drew, P. R., & Dewe, M. D. (1992). Special collection management. *Library Management*, *13*(6), 8–14.

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., Jancke, G., Sarin, R., & Robbins, D. (2003). Stuff I've seen: A system for personal information retrieval and re-use. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 72–79.

Ellis, D., & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, *53*(4), 384–403.

Elsweiler, D., Baillie, M., & Ruthven, I. (2008). Exploring memory in email refinding. *ACM Transactions on Information Systems*, *26*(4), 1–36.

Farina, P. A. (2005). *A comparison of two desktop search engines: Google Desktop Search (beta) vs. Windows XP Search Companion*. Paper presented at the 21st Rensselaer at Hartford Computer Science Seminar. Retrieved November 27, 2009, from www.rh.edu/~rhb/cs_seminar_2005/SessionA3/farina

Fertig, S., Freeman, E., & Gelernter, D. (1996a). Finding and reminding reconsidered. *SIGCHI Bulletin*, *28*(1), 66–69.

Fertig, S., Freeman, E., & Gelernter, D. (1996b). Lifestreams: An alternative to the desktop metaphor. In M. J. Tauber (Ed.), *Conference companion on human factors in computing systems: Common ground* (410–411). New York: ACM Press.

Fisher, D., Brush, A. J., Gleave E., & Smith, M. (2006). Revisiting Whittaker & Sidner's "Email Overload": Ten years later. *Proceedings of the 20th Anniversary ACM Conference on Computer Supported Cooperative Work*, 309–312.

Frohlich, D., Kuchinsky A., Pering C., Don, A., & Ariss, S. (2002). Requirements for photoware. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 166–175.

Gilbert, D. (2006). *Stumbling on happiness*. New York: Knopf.

Golder, S., & Huberman, B. (2006). The structure of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198–208.

Gonçalves, D., & Jorge, J. A. (2003). An empirical study of personal document spaces. *Proceedings of the International Workshop on Design Specification, and Verification of Interactive Systems*, 46–60.

Gonçalves, D., & Jorge, J. A. (2004). Describing documents: What can users tell us?

Proceedings of the International Conference on Intelligent User Interfaces, 247–249.

Gwizdka, J. (2004a). *Cognitive abilities and email interaction: Impacts of interface and task*. Unpublished doctoral dissertation, University of Toronto, Canada.

Gwizdka, J. (2004b). Email task management styles: The cleaners and the keepers. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 1235–1238.

Henderson, S., & Srinivasan, A. (2009). An empirical analysis of personal digital document structures. *HCI International*, 394–403.

Hearst, M. A. (1999). User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern information retrieval* (pp. 257–322). Boston: Addison-Wesley.

Jones, W. (2004). Finders, keepers? The present and future perfect in support of personal information management. *First Monday*, 9(3). Retrieved April 3, 2010, from www.firstmonday.dk/issues/issue9_3/jones/index.html

Jones, W. (2007a). *Keeping found things found: The study and practice of personal information management*. San Francisco, CA: Morgan Kaufmann.

Jones, W. (2007b). Personal information management. *Annual Review of Information Science and Technology*, *41*, 453–504.

Jones, W., Bruce, H., & Dumais, S. (2003). How do people get back to information on the Web? How can they do it better? *Proceedings of the International Conference on Human-Computer Interaction*, 793–796.

Jones, W., & Dumais, S. (1986). The spatial metaphor for user interfaces: Experimental

tests of reference by location versus name. *ACM Transactions on Office Information Systems*, *4*(1), 42–63.

Jones, W., Phuwanartnurak, A. J., Gill, R., & Bruce, H. (2005). Don't take my folders away! Organizing personal information to get things done. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 1505–1508.

Jones, W., & Teevan, J. (2007). *Personal information management*. Seattle: University of Washington Press.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision making under risk. *Econometrica*, *47*, 263–291.

Kalnikaitė, V., Sellen, A., Whittaker, S., & Kirk, D. (2010). Now let me see where I was: Understanding how lifelogs mediate memory. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2045–2054.

Kalnikaitė, V., & Whittaker, S. (2007). Software or wetware? Discovering when and why people use digital prosthetic memory. *Proceedings of ACM SIGCHI Conference on Human Factors in Computing Systems*, 71–80.

Kalnikaitė, V., & Whittaker, S. (2008a). Cueing digital memory: How and why do digital notes help us remember? *Proceedings of the British Computer Society Conference on Human Computer Interaction*, 153–161.

Kalnikaitė, V., & Whittaker, S. (2008b). Social summarization: Does social feedback improve access to speech data? *Proceedings of ACM Conference on Computer Supported Co-operative Work*, 9–12.

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

Kelly, D., & Teevan, J. (2007). Understanding what works: Evaluating personal information management tools. In W. Jones & J. Teevan (Eds.), *Personal information management* (pp. 190–205). Seattle: University of Washington Press.

Kidd, A. (1994). The marks are on the knowledge worker. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 186–191.

Kirk, D., Sellen, A., Rother, C., & Wood, K. (2006). Understanding "photowork". *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 761–770.

Klimt, B., & Yang, Y. (2004, July). *Introducing the Enron corpus*. Paper presented at the First Conference on Email and Anti-Spam, Mountain View, CA. Retrieved April 3, 2010, from www.ceas.cc/papers-2004/168.pdf

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the Web. *ACM Computing Surveys*, *32*(2), 144–173.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, *42*(5), 361–371.

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

Lansdale, M., & Edmonds, E. (1992). Using memory for events in the design of personal filing systems. *International Journal of Man-Machine Studies*, *36*, 97–126.

Lowe, M. (2006). *Evaluation of desktop search applications* (Technical report). Kalio: Sydney, Australia.

Mackay, W. E. (1988). More than just a communication system: Diversity in the use of

electronic mail. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, 344–353.

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

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge, UK: Cambridge University Press.

Marshall, C., (2008a). Rethinking personal digital archiving, Part 1: Four challenges from the field. *DLib Magazine*, *14*(3/4). Retrieved April 29, 2010, from www.dlib.org/dlib/march08/marshall/03marshall-pt1.html

Marshall, C., (2008b). Rethinking personal digital archiving, Part 2: Implications for services, applications, and institutions. D-Lib Magazine, 14(3/4). Retreived April 29, 2010, from www.dlib.org/dlib/march08/marshall/03marshall-pt2.html

Millen, D., Yeng., M., Whittaker, S., & Feinberg, J. (2007). Social bookmarking and exploratory search. *Proceedings of the European Conference on Computer Supported Co-operative Work*, 179–198.

Morris, D., Ringel Morris, M., & Venolia, G. (2008). SearchBar: A search-centric Web history for task resumption and information re-finding. *Proceeding of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 1207–1216.

Nardi, B., Whittaker, S., Isaacs, E., Creech, M., Johnson, J., & Hainsworth, J. (2002, April). ContactMap: Integrating communication and information through visualizing personal social networks. *Communications of the ACM*, *45*(4), 89–95.

Obendorf, H., Weinreich, H., Herder, E., & Mayer, M. (2007). Web page revisitation

revisited: Implications of a long-term click-stream study of browser usage. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 597–606.

Osburn, C. B., & Atkinson, R. (1991). *Collection management: A new treatise*. Greenwich, CT: JAI Press.

Pazzani, M. J. (2000). Representation of electronic mail filtering profiles: A user study. *Proceedings of the International Conference on Intelligent Use Interfaces*, 202–206.

Petrelli, D., Whittaker, S., & Brockmeier, J. (2008). Autotopography: What can physical mementos tell us about digital memories? *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 53–62.

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

Pirolli, P., & Card, S. K. (1995). Information foraging in information access environments. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 51–58.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review*, *106*, 643–675.

Ringel, M., Cutrell, E., Dumais, S., & Horvitz, E. (2003). Milestones in time: The value of landmarks in retrieving information from personal stores. *Proceedings of Human-Computer Interaction (INTERACT '03)*, 184–191.

Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., & van Dantzich, M. (1998). Data mountain: Using spatial memory for document management.

Proceedings of the ACM Symposium on User Interface Software and Technology, 153–162.

Rodden, K., & Wood, K. (2003). How do people manage their digital photographs? *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 409–416.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Russell, D., & Lawrence, S. (2007). Search everything. In W. Jones & J. Teevan (Eds.), *Personal information management* (pp. 153–166). Seattle: University of Washington Press.

Sellen, A., & Harper, R. (2002). *The myth of the paperless office*. Cambridge, MA: MIT Press.

Shannon, C., & Weaver, W. (1949). *A mathematical theory of communication*. Urbana: University of Illinois Press.

Tang, J. C., Lin, J., Pierce, J., Whittaker, S., & Drews, C. (2007). Recent shortcuts: Using recent interactions to support shared activities. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 1263–1272.

Tang, J. C., Wilcox, E., Cerruti, J. A., Badenes, H., Nusser, S., & Schoudt, J. (2008). Tag-it, snag-it, or bag-it: Combining tags, threads, and folders in e-mail. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2179–2194.

Tauscher, L., & Greenberg, S. (1997). How people revisit Web pages: Empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, *47*(1), 97–137.

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. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 415–422.

Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Venolia, G., Gupta, A., Cadiz, J. J., & Dabbish, L. (2001). *Supporting email workflow* (MSR-TR-2001-88). Redmond, WA: Microsoft Research.

Venolia, G., & Neustaedter, C. (2003). Understanding sequence and reply relationships within email conversations: A mixed-model visualization. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 361–368.

Wagenaar, W. (1986). My memory: A study of autobiographical memory after six years. *Cognitive Psychology*, *18*, 225–252.

Wattenberg, M., Rohall, S., Gruen, D., & Kerr B. (2005). Email research: Targeting the enterprise. *Human Computer Interaction*, *20*(1–2), 139–62.

Wen, J. (2003). Post-valued recall Web pages: User disorientation hits the big time. *IT & Society*, *1*(3), 184–194.

Whittaker, S. (2005). Supporting collaborative task management in email. *Human-Computer Interaction*, *20*(1–2), 49–88.

Whittaker, S., Bellotti, V., & Gwizdka, J. (2007). Everything through email. In W. Jones

& J. Teevan (Eds.), *Personal information management* (pp. 167–189). Seattle: University of Washington Press.

Whittaker, S., Bergman, O., & Clough, P. (2010). Easy on that trigger dad: A study of long term family photo retrieval. *Personal and Ubiquitous Computing*, *14*(1), 31–43.

Whittaker, S., & Hirschberg, J. (2001). The character, value and management of personal paper archives. *ACM Transactions on Computer-Human Interaction*, *8*(2), 150–170.

Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., et al. (2002). SCANMail: A voicemail interface that makes speech browsable, readable and searchable. *Proceedings of the ACM Conference on Human Computer Interaction*, 275–282.

Whittaker, S., Jones, Q., & Terveen, L. (2002a). Contact management: Identifying contacts to support long term communication. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 216–225.

Whittaker, S., Jones, Q., & Terveen, L. (2002b). Managing long term conversations: Communication and contact management. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*. Retrieved May 7, 2010, from www.computer.org/portal/web/csdl/proceedings/h#4

Whittaker, S., Jones, Q., Nardi, B., Creech, M., Terveen, L. Isaacs, E., et al. (2004). Contactmap: Organizing communication in a social desktop. *ACM Transactions on Computer-Human Interaction*, *11*(4), 445–471.

Whittaker, S., & Sidner, C. (1996). Email overload: Exploring personal information

management of email. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 276–283.

Whittaker, S., Terveen, L., & Nardi, B. A. (2000). Let's stop pushing the envelope and start addressing it: A reference task agenda for HCI. *Human Computer Interaction*, *15*, 75–106.

Wilhelm, A., Takhteyev, Y., Sarvas, R., Van House, N., & Davis, M. (2004). Photo annotation on a camera phone. *Extended Abstracts on Human Factors in Computing Systems, CHI '04*, 1403–1406.

Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation*, *37*(1), 3–15.

Wilson, T. D. (1994). Information needs and uses: Fifty years of progress? In B. C. Vickery (Ed.), *Fifty years of information progress: A Journal of Documentation review* (pp. 15–51). London: Aslib.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation*, *55*(3), 249–270.