Personal Information Management: From Information Consumption to Curation

Steve Whittaker

IBM Research Almaden

ABSTRACT

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 challenge this *consumer* characterisation 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 a *personal* resource that we *keep, manage* and (sometimes repeatedly) *exploit*. I call this information *curation*. I first summarise arguments against 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, namely documents, emails, 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 there are surprising benefits for strategies that rely on piles as opposed to files. And despite the emergence of desktop search, exploitation currently remains reliant on manual methods such as navigation. There are a number of new technologies that could potentially address important curation problems, but implementing these in acceptable ways remains a challenge. I conclude with a summary of outstanding research and technical questions.

1. 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 and Weaver, 1949). Consistent with this, a new set of computer science theories argue that our general information behaviour is akin to the *foraging* behaviours of hunter gatherer peoples (Pirolli, 2007, Pirolli and Card, 1995, 1999). According to this view, people actively *seek out* and *consume new* information from public resources. While not as extreme in its focus on consumption, the information science literature also emphasises *discovery* of new public information, rather than its exploitation. Many different models have been proposed to characterise how people find new information in public collections (Belkin, 1980, Ellis and 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 organised or curated for future use.

For example, in a very influential theory, Belkin (1980) proposes that people are motivated by ASK (an anomalous state of knowledge) to discover new relevant information. He talks about the steps

that people follow to address that anomalous state. Kuhlthau (1991) describes various information seeking processes, including recognising an information need, and identifying a general topic, as well as stages for formulating and gathering information. A similar *feature set model* is proposed by Ellis and Haugan (1997) who detail the activities involved in finding information, including browsing, chaining, monitoring, differentiating, extracting and verifying information. Wilson (1999) provides a high level macro-model which characterises how information needs arise, and what aids or hinders these processes of information seeking, incorporating insights from these 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 only about how public information is *found*, and ignore what happens after finding has occurred. In a systematic meta-analysis of theoretical information science, Wilson (1999) confirms that information science theories have not tackled what he calls 'information use', i.e. 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 behaviours which I call *personal information curation.* While there are some studies of curation behaviours in the collection management literature, these have tended to focus on the activities of information professionals who are trained to organise and manage *public* collections (Drew and Dewe, 1992, Osburn and Atkinson, 1991). Similar studies within computer supported co-operative work look at how teams self-organise to create shared repositories (Ackerman 1998, Ackerman and Halverson, 2004, Berlin et al., 1993). In both these cases, however, the focus of curation is on organisation of *public* and not personal collections. Here I will review evidence showing that people's everyday information habits are frequently focused around managing personal data and don't 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 behaviours from my own research and that of my collaborators.

1.1 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 that people persistently engage in active and extensive *preservation* and *curation* behaviours in their information environments. Much as we might want to, we don't immediately delete each email we receive, once we have read or replied to it. And after creating a document or presentation, we don't 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 et al., (2007) summarise 8 studies of *email,* showing that people archive a huge number of messages, with an average of around 2846 messages being kept. Unsurprisingly, these personal email archives are growing larger, with more recent studies (Fisher et al., 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 2200 personal files stored on people's hard drives. And a recent study of digital photos found an average of over 4000 personal pictures (Whittaker et al., 2010). Studies of web bookmarking show that people also preserve hundreds of bookmarks (Abrams et al., 1998, Aula et al., 2005, Boardman and Sasse, 2004, Cockburn and Greenberg, 2000, Catledge and Pitkow, 1995). And of course these behaviours aren't 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 62kg of paper, equivalent to a pile of phone directories 30m high.

Furthermore it's not just that people passively *keep* this information, they also make strenuous attempts to *organise it in ways that will promote future retrieval.* For email, Bellotti et al (2005) found that people spend *10% of their total time in email* filing messages, leading to an average of 244 folders in their email collections. Personal computer files are organised in a similar way, with people averaging 57 folders with an average depth of 3.3 subfolders (Boardman and Sasse, 2004). Studies of web bookmarking also show active organisational efforts leading to an average of 17 folders with complex subfolder structure (Abrams et al., 1998, Aula et al., 2005). And Marshall (2008a,b) describes the arcane organisations that result from attempts to preserve information over many years.

So while it's 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 behaviour 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 storage mean that people's hard drives are now filling up with huge amounts of personal photos, videos and music (Bell and Gemmell, 2009, Kalnikaite et al., 2010, Marshall, 2008a,b).

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 aren't seeking out *novel* information. Accessing public the web usually entails *reaccessing previously visited resources*. Various studies have shown that most of people's web behaviour concerns *reaccess*, i.e. returning to information they have already looked at. Between 58 and 81% of all user web accesses are of pages that the user *has accessed previously* (Cockburn and Greenberg, 2000, Obendorf et al., 2007, Tauscher and 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 wasn't possible to create and keep significant personal digital archives. However the prevalence of keeping and re-use suggest a need to develop theories of *curation*, i.e. 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. While other work has neglected how we acquire and manage personal information, one exception is the work of

Jones and colleagues (Bruce et al., 2004, Jones, 2004, 2007, Jones and Teevan, 2007), and we use a variant of Jones' PIM lifecycle framework to organise this review.

The structure of the chapter is as follows. In section 2 we present a framework for the *curation lifecycle,* which describes the processes by which we Keep, Manage and Access information, elaborating the relationships between 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 then review the challenges of Keeping, Management and Exploiting personal information. We present relevant research on how and why people keep information, the different ways they organise it, and finally how they access and exploit that stored information. In each case we review how different types of information (emails, documents, photos, webpages) are treated differently. The final section looks to the future, exploring different technical developments that may influence the future of information curation, as well as outlining outstanding empirical and methodological issues.

2. 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 3 stage model of the curation lifecycle that is a variant of that described in (Jones, 2007, Jones and Teevan, 2007). We talk about the relations between different phases of the lifecycle, and clarify differences between our framework and Jones' work. We also introduce important distinctions between different *properties* of information that have implications for curation behaviours.

2.1 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 anticipate needing in the future. But how do we decide what's worth keeping? What principles govern decisions about the sorts of information we keep (Jones, 2004, 2007)? 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,b)? Keeping is clearly a complex decision that is influenced by many factors, including the *type* of information being evaluated, *when* we anticipate 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, emails, photos or webpages) have different utility and will consequently be processed in very different ways. Transient information encountered on a web page will be treated very differently from a personal document we have been working on for several days, or an email 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 they required careful physical organisation and storage. But the ease of generating pictures may have important downstream consequences for retrieval that need to be taken into account when deciding whether or not to keep them (Whittaker et al., 2010).

2.2 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 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.

There are different ways of managing information that 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, e.g. by filing our incoming information into structured folders. This information should then be more straightforward 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 analysed 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 emails to pile up in our inbox. This tactic reduces the costs of organising this information, but may mean that it is harder to locate critical information when we need it (Malone, 1983, Whittaker and Hirschberg, 2001, Whittaker and Sidner, 1996, Whittaker, 2005).

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 organisation 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 don't 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 out' information that has turned out to be of little value that may be compromising the uptake of information of definite utility. People occasionally work through email inboxes deleting old or irrelevant information (Whittaker and Sidner, 1996). However it is abundantly clear that people find such 'cleanup' activities difficult, not only because they require judgements about the projected value of information, but also there may be emotional investment in information that they have invested time and effort in organising (Bergman et al., 2003, Jones, 2007, Marshall, 2008a, b, Whittaker and Hirschberg, 2001).

2.3 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? Exploitation success clearly relates to keeping and management practices. Careful attempts to organise valuable information should make it easier to reaccess that data. But new technologies potentially reduce the need to organise. Emerging technologies

such as desktop search (Cutrell et al., 2006a, Dumais et al., 2003, Russell and Lawrence, 2007), promise to reduce the overhead of organising our files, because we no longer have to manually navigate to them.

There are two main methods that can be used to exploit information

- 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 information item, one of its attributes or its contents.

There are advantages and disadvantages of both methods. Navigation, being incremental, offers the user feedback at each access stage (Barreau and Nardi, 1995, Bergman et al., 2008), but in the case of complex folder hierarchies can be laborious because of the multiple levels that people have 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.

Relation to Jones' PIM framework

The differences in terminology between our framework and that of Jones and Teevan (2007) and Jones (2007) are shown in Table 1. The frameworks concur in their overall characterisation of key personal information management processes, such as keeping and management. However Jones and Teevan (2007)'s main focus is on finding and refinding of public data such as that found on the web, e.g. 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 email. Where we focus on web data it is in the context of users' efforts to integrate such information into their personal information systems. Our more strict 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 because Jones and Teevan (2007)'s concern is more with public data, 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	**Curation Lifecycle**
Jones, 2007, Jones and Teevan, 2007	
(Re)finding	
Keeping	Keeping
Metalevel activities (managing, maintaining, …)	Management
	Exploitation

Table 1: Contrast between PIM activities proposed by Jones and Teevan (2007) and those used in the current review

2.4 Interrelations Between Keeping, Management and Exploitation

As will be obvious from the above description, there are close relations between the different processes in the curation lifecycle. For example, exploitation success 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, and the more information we keep, the more effort has to go into organising and maintaining that information. More critically, having more information may increase the difficulty of exploitation, as finding information may be harder when there is more information to search.

Past outcomes may also influence future curation behaviours. Past 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 realise that certain types of organisation are less successful in promoting access they may abandon those methods.

2.5 Information Properties

Not all information items are equivalent. We need to distinguish between different *information properties,* as these differences have implications for the ways in which each type of item will be curated.

Informative versus Action-oriented items

Compare, for example, an average email message, with a page found in a web search. One crucial property of many email 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 emails 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.* While the page may be diverting, 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 email is action oriented (e.g. when people send us FYIs) and not every web page 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 email folders, or leaving active documents on the desktop), have to be set up to prompt the user 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 mustn't 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.

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. Those who have lost data following system crashes are only too aware that if we do not take responsibility for storing and maintaining unique data, then it will not be preserved for future access (Marshall, 2008a,b). 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. Email data lies 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 can't guarantee they will have kept that information.

It is important to note that uniqueness is defined *subjectively:* relative to our own goals and interests. There are innumerable unique information items in the world, but as curators we are only concerned 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 why 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 in 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. No-one 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 organised. The same is largely true for emails: we need to have reliable methods for reaccessing these because we cannot always rely on others to keep the important messages that we need. Web pages 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.

Information Type	**Action vs. 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
Email	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.

Table 2 shows the key properties of common classes of information, such as paper, electronic files, email, photos and web documents. The Table shows these are very different with respect to action-orientation and also for uniqueness. Current paper and electronic documents and emails 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 emails, and personal documents in such a way that promotes effective action.

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 between these processes. In what follows, we review each of these processes separately, but we should not lose sight of the relationships between them.

3. KEEPING

3.1 OVERVIEW, PROBLEMS AND STRATEGIES

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

- Management costs – we need to organise 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 organised 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 organisation.

- 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.

Keeping decisions are a fact of life. Every day we receive new emails, create new files and folders and browse new web sites. 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. Data extracted from Boardman and Sasse (2004) suggests that users acquire an average of 5 new files per day, 5 emails[1] per day, and one bookmark every 5 days. Other studies indicate people acquire one new contact per day (Whittaker et al., 2002a), and around 5 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 emails per day; so focusing exclusively on what they actively decide to keep overlooks the 37 decisions they make to delete irrelevant information. For email alone, making the highly

[1] Although Dabbish et al. (2005) suggest higher keeping rates for email.

conservative assumption that email 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 et al., 2009, Boardman and Sasse, 2004, Jones, 2004, 2007, Whittaker and Hirschberg, 2001, Whittaker and 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. There is a great deal of psychological research that shows that people are poor at making many types of decisions that involve their future. Such prediction requires people to reason about hypothetical situations, which they are notoriously poor at. 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 and Tversky, 1979).

In the PIM context, the keeping decision requires people to predict future informational contexts and assess future informational requirements. In deciding what to keep, people have to evaluate the potential future utility of keeping an item, and weigh this against potential management and exploitation costs associated with keeping it. Jones (2004) argues that the decision whether "to keep or not to keep" information for future usage is prone to two types of costly mistakes. On the 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 and Sasse, 2004, Whittaker and 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 and Gelade, 1980). Furthermore, there is a "deletion paradox": while 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).

When people weigh up 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 and Hirschberg, 2001). However there are also less rational reasons why people avoid deletion, which can be attributed to general psychological decision making processes (Kahneman and 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. And 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 people's keeping decisions, when managing their paper archives, emails, contacts, web pages and personal photos.

3.2 KEEPING PAPER

Somewhat curiously, despite the prevalence of keeping decisions, there have been relatively few studies that have looked directly at this. One exception is a study of people's paper archiving behaviour

(Whittaker and Hirschberg, 2001). While 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 behaviour 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 make decisions 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 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 when we did, we capitalised 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.

Discarding Behaviour

People experienced major problems in deciding what to keep and what to throw away. As the psychology literature would suggest, there was a bias towards 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 rationalising their data, and reported that this was a difficult process. Despite these efforts, the final amount of information that people actually threw away was small compared with what they kept: people discarded just 22% of their original archives, with the final preserved archive on average being more than 18 mover's boxes (equivalent in volume to a pile of telephone directories about 30m high).

We looked at the characteristics not only of what people kept, but also what was discarded. As we expected, at least part of discarded data was once valuable information that had become *obsolete*. As jobs, personal interests or company strategy changes, then the value of particular information decreases. But not all discarded information underwent the transition from valuable to obsolete. For example, 23% of discarded data was *unread*. Why would people keep information that they had never even looked at? 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 acquire large amounts of data that later turn out to be extraneous.

Information overload refers to the fact that people have insufficient time to process all the information they are exposed to. 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 rationalising ("clean-up") of archives. Consequently, people seldom discover that their unread non-urgent documents are superfluous until exceptional circumstances (like the current office move) force people to scrutinise their archives.

Yet even when people find the time to systematically examine new information, 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 judgements 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 rationalise them, so their archives end up containing considerable amounts of information of dubious value. Thus, 74% of our users had not cleaned out their archives for over a year. Furthermore, very few clean-ups occur spontaneously: 84% 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 illustrates important aspects of keeping. When extraordinary events such an office move occur, then people discard about 22% of their data, some of which is obsolete. However other factors besides 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'.

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 that people kept would be unique to that person; because other people will not take responsibility for retaining highly personal data. In contrast, we expected people to be much less likely to keep publicly available data. Why take responsibility for data that is 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% 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% of people's original archive was unique: 15% was unread, but 36% 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 didn't *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 (Kalnikaite and Whittaker, 2007, 2008, Sellen and Harper, 2002). Although most people made such annotations, they seemed of little long-term use, however. 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 (Kalnikaite and Whittaker, 2007, 2008).

3.3 KEEPING EMAIL

Email 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 email is *actionable*, i.e. 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 an action. Another significant contrast to self-created files is that most emails 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 email information. A final rather different characteristic of email is its sheer variability. In our inboxes we may see many different types of messages including: tasks or todo items, documents/attachments, fyis, schedules, social messages and jokes. Again this heterogeneity makes the keeping decision rather different from other information types.

Overall we keep about 70% of our emails (Dabbish et al., 2005). This seems a surprisingly high retention rate given the apparent irrelevance of many of the emails we receive, but there are reasons for this. In what follows, we separately discuss people's keeping behaviours for informational versus actionable messages, as keeping behaviour is very different for each of these.

Informational messages

Informational messages form about one third (34%) of what is delivered in email (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 emails are 'kept around' allowing more informed judgements 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: and 28% 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' as such folders are seldom revisited (Whittaker and Sidner, 1996). A more common strategy is to leave them in the inbox: Dabbish et al. found 42% of informational messages are left in the inbox, to increase 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 re-visited as users reaccess 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 et al., 2003, Whittaker, 2005, Whittaker and Sidner, 1996).

As with paper archives, people experience information overload in email. 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 and Sidner, 1996). One factor contributing to whether a message is read or not is its length and Whittaker and Sidner (1996) found that 21% of inbox messages contained more than 5 screenfulls of text, consistent with the fact that people leave longer messages there for later reading.

Actionable Items

Actionable messages are those that we have to do something specific about. 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 don't forget to deal with it, and they don't 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% of actionable messages immediately (Dabbish et al., 2005). An immediate reply clearly reduces the chance that they will forget to act on a message. However even when people do reply immediately *they still keep 85% of actionable messages*, suggesting that one touch does not describe actual practice.

There are several reasons for such retention. In some cases, one touch and an immediate reply are not possible. Many important email tasks are too complex or lengthy be executed immediately (Bellotti et al., 2005, Venolia at al., 2001, Whittaker and Sidner, 1996, Whittaker, 2005). This leads to deferral of 37% of actionable messages (Dabbish et al., 2005). Deferral is often a direct consequence of *interdependent* tasks, i.e. those collaborative tasks 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, 2003, Whittaker and Sidner, 1996). People may need to negotiate exactly what a collaborative email 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 part of the task. One way to estimate the prevalence of interdependent tasks is by determining how many emails are part of a conversational thread, as

threads indicate relations and common underlying activities among messages. Threading estimates range from 30-62% of messages (Bellotti et al., 2003, Whittaker et al., 2007).

The need for deferral of 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% are deleted). This figure is much higher than for information messages, 30% of which are deleted. Furthermore, actionable messages have to be kept in a way that *guarantees that they will be reencountered*. It's no good deferring 'todo' emails, 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.

3.4 KEEPING CONTACTS

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

It is complex to decide on important contacts from the many people that you are exposed to on a daily basis. As with paper and email archives, it is hard to anticipate 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 despite this strategy, participants were exposed to many more contacts than they recorded information about.

We identified specific factors that were critical in determining important contacts. Just as with deferred evaluation in email 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 email archives, and asked them to distinguish between important and unimportant ones. The findings were quite striking. Despite having huge archives of contacts (858 on average), participants rated only 14% (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 emails. They also excluded spammers.

Overall there are interesting parallels between contacts, paper and emails. People are exposed to many more contacts than they can record systematic information for, so they defer making decisions managing to reserve judgment and 'overkeep' data about contacts that they don't need. Furthermore, the criteria that people use to judge the value of contacts are based around usage and interaction: valued contacts are those who are interacted with often, frequently and recently. However one key difference between contacts, email and papers is that users ignore or 'discard' a much higher percentage of encountered contacts.

3.5 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; e.g. people expend energy creating bookmarks that they never subsequently use. Tauscher and Greenberg (1997) showed that 58% of bookmarks people are never used, suggesting poor decision making.

At the same time, other studies of web behaviours reveal failures of *omission* - where people don't preserve information that *does* turn 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 realising their utility. His study showed that people were only able to later find about 20% of information they have previously accessed and attended to, in an earlier information retrieval session. Such failure partially originates 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' up their current bookmark collection. These findings were replicated in other similar studies (Aula et al., 2005). Instead users preferred to try to retrace their original searches – a strategy which is often unsuccessful.

3.6 Keeping Photos

With the advent of digital photography, the numbers of pictures that people are now taking has increased massively (Bentley et al., 2006, Kirk et al., 2007, Whittaker et al., 2010, Wilhelm et al., 2004), and similar keeping issues are beginning to arise for digital photos. Now that people have collections of thousands of digital pictures, how do they decide which of these to keep and which to delete?

We looked at this in a study of parents with young families (Whittaker et al., 2010) who had an average of 4475 digital pictures. All participants deleted some pictures, both when pictures were taken, and when they were uploading from picture to camera. Participants estimated they deleted on average 17% of their pictures. The reasons people gave for deletion were that the pictures were poor technical quality or did not capture the 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., 2007), 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 weren't therefore focused on the exploitation/retrieval context when they made keeping decisions. As with paper and email, when we probed people further about this conservative approach to keeping, people had a strong expectation that they would return to their photo collection to rationalise it at a later date. And as in our paper and email studies, this rationalisation seldom occurred.

3.7 KEEPING SUMMARY

1. Keeping decisions are difficult because they require people to: (a) predict their future retrieval needs, (b) take into account the possibility that those information needs may change, and (c) make utility decisions under conditions of information overload, often on incomplete readings of information.
2. Errors are made: the primary tendency is overkeeping, i.e. keeping things that are never accessed (observed with paper, email, contacts and photo archives), although there is evidence from some web studies of failing to keep information that later turns out to be relevant.
3. Consistent with overkeeping, deletion is relatively infrequent, varying between around 17% for photos to 30% for emails. Contacts are very different, however, it seems that because people are exposed to many more of these, they are happy to ignore 86% of the contacts they encounter.
4. The nature of the information item affects the keeping decision. This decision is relatively straightforward for certain items: we obviously need to keep unaddressed actionable emails or unique personally generated items that no-one else will safeguard. However it's very hard for people to decide the value of data such as public web pages or informational emails.
5. Because of problems in making the keeping decision, rather than viewing keeping as a one-time decision, people often used a *deferral strategy* – waiting to see whether information is useful. Two major weaknesses of deferral are (a) that people seldom return to their collections to carry out a re-evaluation of tentatively kept information; (b) deferral means that collections are full of items of dubious value - that make it more difficult to find truly valuable information.
6. People don't generally seem to be aware of the implications of overkeeping. While they complain about how full their inboxes are, they nevertheless delete only 30% of emails, and even after spending days working though paper archives they still preserve 78% of those. On the web, in contrast, there is a suggestion that people don't 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.

4. MANAGEMENT

4.1 OVERVIEW, PROBLEMS AND STRATEGIES

We will first describe different methods for organising information, as well as the trade-offs between these. We next discuss factors which influence users' choice of management strategies and studies evaluating the utility of these different strategies. We then briefly talk about a radical alternative which proposes that we forgo preparatory organisation 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 clearly accumulate (Marshall, 2008a,b). Using current estimates of how many documents, digital photos and emails we acquire on a daily basis (Boardman and 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 thousand documents, 115 thousand emails and 120 thousand digital photographs. How

people organise and maintain this information will obviously have a strong bearing on their success in exploiting that information in the future.

Certain types of management also take place more often than we might expect. For certain items such as files and emails, people are perpetually and actively engaged in re-organisation, as reflected by the frequent small modifications they make to their information. For example a longitudinal study (Boardman and Sasse, 2004) found that people create a new file folder *every three days* and they make a new email folder *every 5 days*. In each case, the new structure reflects the fact that people are constantly reflecting on how their information is currently organised and finding it to be inadequate. However, as we mentioned in the keeping section, people seldom engage in major reorganisations or extensive deletion. Instead they tend to incrementally modify existing structures. They are highly unlikely however to monitor and re-organise photos or contacts, for reasons that will become clear.

People also make management mistakes. They often engage in *counterproductive* behaviours in organising their information. Studies of web bookmarking show people construct complex hierarchical bookmarking systems (Abrams et al., 1998, Aula, 2005). Yet we have already seen that users *never* access 42% of the bookmarks they organise for later retrieval (Tauscher and Greenberg, 1997). Efforts organising emails may also not bear fruit. Email filing accounts for 10% of total time in email (Bellotti et al., 2005), and yet information is usually accessed by browsing the inbox or search, rather than folder access (Whittaker, 2005, Whittaker et al., 2007, Tang et al., 2008). With personal photos they may make the opposite type of mistake and fail to organise 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 organisation - in part because people believed that they would be able to retrieve their photos without needing to organise them (Whittaker et al., 2010).

Semantic organisation

Organising information is a fundamental cognitive activity. One basic approach is to *apply conceptual organisation* to information. Even newborn infants categorise objects, with natural psychological categories tending to be based around exemplars or prototypes. For example, people's concept of 'bird' is based around exemplars such as robins, rather than unusual cases such as penguins. Our judgements and reasoning are influenced by the extent to which particular instances are similar to those exemplars (Rosch et al., 1976, Rosch, 1978).

When managing personal information, there are two different and separate aspects to organisation that are important for effective exploitation. We call these *mental* and *physical* cueing. As many psychological studies have shown, the mental act of imposing organisation on information makes it inherently more memorable. Organising things within a consistent conceptual structure means that, at recall, one item may mentally trigger memory of a related one, so applying semantic organisation is highly effective in promoting recall (Baddeley, 1995, Craik and Lockhart, 1972). Organisation helps recall, even if people don't have direct access to their organisational scheme at retrieval. For example in a recent study we showed that the simple act of organising conversational information by taking notes increased recall *even when people didn't use their notes at retrieval time* (Kalnikaite and Whittaker, 2008). Organisation is also important because the *products* of organisational efforts can themselves be used as *physical* retrieval cues. Appropriate notes can serve as cues that remind us about information items that we might otherwise have forgotten (Kalnikaite and Whittaker, 2007, 2008). Well chosen folder names serve to cue

people about their contents and organisation (Bergman et al., 2003, Lansdale, 1988, Jones and Dumais, 1986, Jones et al., 2005).

Organisation and labelling are in the mainstays of most computer operating systems. The main way that people organise their digital information is to recursively sort it into categories (in directories, folders or subfolders) and then apply meaningful labels to these folders and subfolders. The act of applying organisation may help retrieval by mental cueing, as well as generating a navigable conceptual structure with folder labels serving as physical retrieval cues. Note also that folders usually contain a strong spatial component – with subfolders 'sitting inside' super-ordinate items, and this too can help cue retrieval (Jones and Dumais, 1986).

Temporal organisation and reminding There is a second, less obvious, type of organisation that has been less extensively researched. We have already seen that some important information that people deal with is *actionable.* Further it is usually the case that those actions are *required to happen by a certain time*, e.g. to meet a certain deadline. People must therefore ensure that *actionable* information is organised 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 organisational structure allowing access to any item, if you forget the deadline relating to that information. Reminding will turn out to be a critical problem in what follows, especially in the case of email, where actionable items are prevalent.

Most psychology research on organisation has looked at *natural categories* (e.g. how we mentally organise places, events, names and faces). But it has not looked at the types of information we are addressing here, namely *synthetic,* human-generated information such as documents, emails, photos or web pages. Nevertheless there is considerable HCI and library science research looking into people's preferences for organising such personal data. For example, people prefer to relocate their documents *spatially* rather than using keyword search (Barreau and Nardi 1995, Bergman et al., 2008). This spatial organisation works even better when the document space is three-dimensional although this may not scale well to large number of files (Robertson et al. 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 and 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 and Simpson, 1990). There is also evidence for the utility of temporal organisation 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 et al., 2003). The importance of temporal factors is also shown by log files of search tools revealing a bias towards retrieval of *highly recent* information (Dumais et al. 2003; Cutrell et al. 2006a).

In addition to these overall organisational preferences, other work has explored different types of management strategy and what motivates people to choose them. We now describe strategies for paper, digital files, email, 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 organisation of our personal data will soon become obsolete. Improvements in desktop search will mean that documents emails and web pages can be easily retrieved without the need for active organisation (Russell and Lawrence, 2007, Cutrell et al., 2006a). This is an appealing idea. We have seen that management activities are onerous and difficult for

users, who may invest in organisational efforts that are not always directly successful. We will discuss these claims in more detail when we discuss exploitation techniques and evaluate the efficacy of different search tools.

4.2 Managing Paper

Malone (1983) conducted a pioneering study into people's organisational 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 organisational strategies. Piles are easier to create and maintain, as they are less systematic. They have a less clear organisational structure with more items in each pile, which may make retrieval within each pile more inefficient. But because there tend to be fewer piles in total, this leaves fewer potential locations to be searched which may compensate for this lack of organisation. 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 above, we investigated when and why people choose filing or piling strategies. The distinction between filers and pilers was not absolute, instead being 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%.

Pilers often amass information without attempting to systematically organise it. This laissez faire approach should lead to an accumulation of unscrutinised information before the office move. However 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, which turns out to be of little utility, that they later have to discard. 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 organised systems, we expected filers to find it easier to find data, and they should 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 potentially located in a relatively small number of piles that people frequently sift through. 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 the pile. 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 rationalising complex organisations of documents. Similar findings are reported in a study comparing folders and tags as methods of organising personal information (Civan et al., 2008).

A final possible reason why 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 in consequence they are able to access less of it. Consistent with this is the fact that the absolute amounts of data accessed by both groups were very similar.

We also expected filers to be quicker to rationalise their data in preparing for the move, given the greater care they have taken to initially organise their data. But there were no differences in packing time for filers and pilers. This could be because pilers' greater organization is offset by having more data to sift through. And contrary to our predictions, pilers found it subjectively easier to rationalise archives in preparation for the move. Why was this? Despite the fact that 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 organising. 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.

4.3 Managing Digital Files and Folders

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

many items unfiled (in a manner similar to paper piling), with a small proportion (6%) leaving most items unfiled. In some cases, people did not file actionable documents (i.e. those that they were currently working on), 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 2 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 & 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 found to be 8.45, with an average branching factor (which is an estimation of the mean number of subfolders per folder) of 1.84 indicating a deep and narrow hierarchy structure. 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 & Jorge, 2003) and 11.1 for (Henderson & Srinivasan, 2009).

In another study probing the reasons why people generate specific folder structures, Jones et al. (2005) interviewed people about the nature of their folder systems. They discovered that, consistent with physical cueing, many folders were seen as *plans,* i.e. structures that people used to organise their future work. They found that 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 used also various workarounds to make various types of information more salient, e.g. by labelling folders 'aacurrent' instead of 'current' to ensure that this information was more obvious when browsing an alphabetically ordered folder list.

Bergman et al. (2003, 2009) also document workarounds *within* folders, to make individual files and folders more salient, at the same time 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 organisation is the *adaptive* nature of active folders. Bergman et al (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/day. One implication of this continual re-access is that users are likely to discover suboptimal organisation, leading them to adaptively 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 turn out 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 3 days. In contrast, with digital pictures, months may elapse between new folders being created, with negative effects of people's ability to retrieve those pictures (Whittaker et al., 2010).

More recently, new types of tool have been developed to support different types of organisation. One example is *tagging*. Phlat (Cutrell et al., 2006b) is a system that 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 (as 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 terms of the ways that data can be accessed and navigated to. 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 that data. But despite 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 enough of the tags needed 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 these costs of creating personal tags.

4.4 Managing Email

Actionable Items

Managing email is complex, and different from paper or standard digital files. A critical aspect of email is that it contains many *actionable* messages. To be effective, people need to organise actionable information in such a way that they are reminded of what they need to do when. This means that users have to organise 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, and most actionable items must wait to be processed. Dabbish et al. (2005) found that on average 37% of messages that require replies are deferred, which equates to about 4 deferred messages per day. If people forget these deferred tasks, this can create major headaches both for the user and their organisation.

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 hopefully be reminded about their outstanding actionable messages. Dabbish et al. too report that actionable items are left in the inbox around 79% of 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% users had experimented with a strategy of filing actionable items in a "todo", folder. In fully 95% of these cases, this folder was abandoned, because people had to explicitly remember to go to it, open it and review its contents. This extra effort contrasts with being reminded about outstanding actions by the mere fact of seeing them in the inbox when reading new email. Although other studies (Bellotti et al., 2003) suggest that some users change their work practices to exploit ‘todo’ folders, this demands extra cognitive steps. Paradoxically, these users have to actively remember 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 the user to continually scroll through their inbox to ensure that

these items are not 'out of sight and out of mind' (Whittaker et al., 2003, Whittaker, 2005, Whittaker and Sidner, 1996). Tang et al. (2008) looked at the proportion of their inbox that users had constantly visible, finding that on average only 25% of inbox emails were in view. The remaining 75% of messages were not therefore serving as direct visual reminders for outstanding actions – compromising their ability to remind.

Other users try to keep their inbox clear by filing incoming actionable items in dedicated task related folders (Bellotti et al., 2005, Whittaker and Sidner, 1996). Whittaker and Sidner dubbed these people '*frequent filers'* and documented how 25% of users create such folders. There are obvious advantages of this strategy: removing such items from the inbox keeps the inbox 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 they file important information and forget about it.

A final strategy for actionable items is a hybrid of the above. Whittaker and Sidner identified a final group accounting for 35% 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 rationalise their inbox. The process would then be repeated with the inbox gradually growing in size until another 'crisis' is experienced and extensive filing takes place once more.

What determines which strategy people choose when processing actionable emails? Whittaker and Sidner (1996) looked at the impact on strategy choice, of organisational role and incoming volume of messages. Managers were more likely to receive greater volumes of email, 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 other research demonstrated relations between cognitive style and strategy (Gwizdka 2004a; 2004b).

Other studies of email have found some support for these management strategies (Whittaker et al., 2002, Whittaker, 2005, Mackay, 1988, Dabbish et al., 2005, Bellotti et al., 2005, Fisher et al., 2008). However later work indicates few instances of pure 'no filers', i.e. people with absolutely no folders who are totally reliant on their inboxes for task management. Balter (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 email they receive increases. He argues that those receiving the highest volumes of email are those with these least time to organise it.

Informational Messages

We now look at how users organise *informational* messages. A substantial percentage of emails are informational as opposed to actionable (Dabbish et al., 2005, Whittaker and Sidner, 1996). Users also experience problems in processing informational emails. Observations of email behaviour show that users spend huge amounts of time overall in organising emails: on average 10% of people's total time in email is spent filing messages (Bellotti et al., 2005).

Again Whittaker and Sidner (1996) examined why users have problems with filing such information. There are several reasons why creating folders for informational messages is hard.

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* their judgment 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. Users may not be able to remember folder labels, especially when users have large numbers of older folders. Research combining multiple studies shows that people have an average of around 39 email 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% of users' folders contain only one or two items. Later studies duplicate these observations although finding a lower percentage (16%) of such ‘failed’ folders (Fisher et al. 2006). Not only do these tiny ‘failed folders’ not significantly reduce the complexity of the inbox, they introduce the dual overheads of: (a) creating folders in the first place, and (b) remembering multiple folder definitions every time there is a decision about filing a new inbox item. This cognitive overhead is illustrated by the fact 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 and Sidner, 1996). Of course a small number of these failed folders may represent new activities that the user is planning to carry out (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 between different 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 and Sidner, 1996).

Elsweiler et al (2008) looked at the impact of filing strategy on users’ memory for their emails. Frequent filers tended to remember less about their emails. This result is consistent with our earlier observations about premature filing. Filing information too quickly can lead to the creation of archives containing spurious information, and quick filing also means that users aren’t exposed to the information frequently in their inbox, making it hard for them to remember its properties or even its existence.

Thus email 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 they may be 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 (6.1).

4.5 Managing Web Pages

Unlike email, web information is largely not actionable: users may want to ensure that they remember to read a webpage, but in general there aren't negative consequences for failing to do this.

One prevalent form of managing web information is to *bookmark* encountered webpages. There have been numerous studies looking into how people organise their bookmarks. Two early studies documented the number of bookmarks created as well as their underlying structure. For example, Abrams et al. (1998) found that 6% of respondents had no bookmarks, 10% had 1-10, 24% had 11-25, 44% had 26-100, 14% had 101-300, and 2% had 300+ bookmarks. And Boardman and Sasse (2004) found that people organised their bookmarks into an average of 17 folders. Another study (Bruce et al, 2004) observed further strategies people use for organising useful web information that they encounter. In addition to bookmarking, users might forward themselves a link in email, 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 et al. 2005 looked at people's bookmark collections finding that 92% have bookmarks, with an average of 220 links, although there is huge variance in collections: 21% of people have fewer than 50 bookmarks, and 6% have none. The largest collection contained 2589 links with 425 folders. Most of Aula et al.'s (2005) informants reported major problems in organising and managing their collections. Consistent with other studies (Tauscher and Greenberg, 1997) users often bookmarked information that they never subsequently revisited. In contrast, other studies showed that users are unwilling to create new bookmarks fearing that creating bookmarks for information of unclear utility will clutter their existing set of useful bookmarks - compromising the utility of useful items (Aula et al., 2005, Wen, 2003). Aula et al 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. There was a subgroup of heavy users of bookmarks, who had collections of over 500 links. These heavy users tended (like email spring cleaners) to clean up their collections from time to time deleting unused or no longer functioning links. They also carefully organised bookmarks into hierarchical levels (similar to a file system). For these users who invested organisational effort, bookmarks seemed to be an indispensable tool. Abrams et al. also looked at the types of strategies that people used for organising their bookmarks. They found 4 main types: about 50% of people were sporadic filers, a further 26% never organised bookmarks into files, around 23% created folders when they accessed a web page and around 7% created folders at the end of a session. Creating folders also seem 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 of the 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 potentially providing richer retrieval cues (Cutrell, 2006b, Lansdale, 1988). More importantly they allow tags to be *shared* between 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 stabilise on common descriptions of web resources so that people can exploit others' tags (Golder and Huberman 2006, Millen et al., 2007). Furthermore, with suitable user interface design, e.g. text completion, problems such as inconsistent spellings can be finessed, as well as promoting 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.

4.6 Managing Photos

Photos are very different from emails and web pages, tending to be *self-generated* (like many files), and being neither informational nor actionable. They are also perceived to be highly important and often irreplaceable (Petrelli et al., 2008, Whittaker et al., 2010). How then do people organise them? Recent studies show that people manage to organize photos using rather rudimentary structures (Kirk et al., 2007, Whittaker et al., 2010).

Whittaker et al. (2010) looked how parents organised family photo archives. They found that these collections tended to have very little hierarchical structure, and were organised 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 storage devices (computers and 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 that folder often contained heterogeneous data, containing pictures that relate to multiple events (possibly because they were uploaded at the same time and never subsequently reorganised).

How can we explain this lack of organisation? Previous work has highlighted how participants are able to exploit their familiarity with *recently taken* pictures to quickly scan, sort and organize materials for sharing with others (Kirk et al., 2007). Possibly as a result of these experiences with recent pictures, our participants may have expected themselves to be very familiar with their *entire* picture collection, and as a result weren't motivated to organize their collections carefully. In most cases, it seemed that people hadn't accessed the vast majority of their pictures since they were uploaded. We saw evidence of this when participants retrieved pictures. Photos almost always appeared in the "list" view. However, participants universally preferred to view pictures in the thumbnail view for easier scanning. Had the participants previously opened these folders, the thumbnail view would have remained at the interview. And because participants seldom accessed pictures, they didn't discover how poorly organised these were. One reason for the lack of organisation 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 potential way to organise might be to *annotate* pictures. However, consistent with earlier studies (Frohlich et al., 2002, Kirk et al., 2007, Rodden and Wood, 2003), we found very little evidence of annotation. One reason for failing to annotate is that this is onerous. Another problem, also

observed in earlier studies (Kirk et al., 2007, Rodden and Wood, 2003), is that users may fail to 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.

4.7 Management Summary

1. Management is a difficult activity for users, because it requires people to predict when or how information will be accessed. To create effective organisation 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.
2. 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 as that workspace can become cluttered with many unrelated actionable items. And one specific problem with using the email inbox for reminding is that as new items arrive they tend to displace older actionable items leading them to be 'out of sight and out of mind'. The disadvantage of dedicated task folders is that these need to be constantly accessed and monitored.
3. 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 their 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.
4. For *informational* items, users experience difficulty in categorising information, failing to accurately predict 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.
5. Both users' dispositions and the volume of information they receive may influence the type of organisational strategy they use. 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 organise their information.
6. Certain types of information such as web pages and photos are infrequently re-accessed. Infrequent access may mean that people fail to realise what information they have available and how poorly organised it is. Tags don't 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.

5. Exploitation

5.1 Overview, Problems and Strategies

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

Exploitation Not Information Seeking.

Exploitation is different from information foraging and classic information seeking. In both foraging (Pirolli et al.., 1999, 2007) 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 (Lansdale, 1988, Bergman et al., 2003). In other words, people are searching their own organisation and not a public database. Second the exploiter may *remember* significant details about the target information item and how it has been organised.

For example, Goncalves and Jorge (2004) asked participants to tell stories about 3 personal documents they had recently worked on. 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 *physical* (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 organisations that promote future exploitation.

Exploitation therefore involves *reconstruction of partially familiar personally organised information,* rather than evaluation of unfamiliar publically organised data. A further difference concerns success criteria: while information seeking, 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 recognise the target document, so they do not have to scrutinise it to determine its relevance as they would an unknown web page. But in other ways access to very specific information can be made harder when access is only satisfied if a specific item is found, and there may be strong feelings of frustration about failure to locate that item (Whittaker et al., 2010).

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, and this requires them to reconstruct important aspects of the target document (e.g. title, keywords, date). If there is a good match between organisational cues and the retrieval context, then retrieval will succeed. But to create effective retrieval cues, users need to successfully anticipate *when* and *how* they will consume information.

There are 4 main ways that we access personal information.

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 to locate our information. People usually manually traverse their organizational hierarchy. 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 key word queries, in the same way they conduct web searches. Users first generate 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 the item was created). The desktop search engine then returns a set of results from which the user selects the relevant item. Search has elsewhere been characterised as a form of *teleporting* whereby users move *directly* to the target information, without the intermediate steps that characterise navigation (Teevan et al., 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 both on the desktop (Cutrell et al., 2006a) or on the web (deli.cio.us, Flickr). They allow users more flexibility in how they categorise the item (as more than one label can be applied. Multiple tags mean richer retrieval cues, as the same information can be accessed via several different tags.

The above 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 et al., 2003, Obendorf et al., 2007). For example users can deliberately bookmark valued information or save it to disk and then navigate back to this data. Or they can apply less effortful strategies such as accessing information via the history list (a list of sites visited), or use the browsers 'back' button to reaccess recent information.

Costs and Benefits of Exploitation Strategies

If the fit between the organisation 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 and Sidner, 1996). Putting information in a folder may decrease its ability to remind which may be vital for actionable information. In addition, because categorisation is itself cognitively challenging, users may create spurious folders that are seldom accessed, and which may make classification of new information harder (Fisher et al., 2006, Whittaker and 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 organisational hierarchy is predictable and includes a spatial component which users find valuable (Barreau and Nardi, 1995, Bergman et al., 2008, Jones and Dumais, 1988, Robertson et al., 1988). 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 accessed the wrong branch of their file hierarchy. At the same time, there are disadvantages to navigation, compared with search. In complex organisational 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 organisation scheme (Hearst, 1999).

Furthermore, users have to remember at retrieval time, *how* information was classified, which can be difficult when there are multiple categorisation possibilities (Lansdale, 1988, Russell and 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 their query any attribute they happen to remember (date, name, filetype) (Lansdale, 1988). Search may also be more efficient: user 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 potentially finesses the *management problem,* as users don't have to apply organisational 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, so that information must be organised in such a way that users encounter it opportunistically. Neither search nor navigation through complex file organisations are appropriate support for actionable items, as both require *deliberate* acts to seek out data, whereas the primary characteristic of actionable items are that these should trigger *automatic reminding.* This is clearly a very hard problem: effective reminding means users don't 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, and it turns out that very different strategies are needed for actionable than informational items.

Turning now to public data (such as web data) that people want to incorporate into their personal organisational schemes, it is apparent that users may have less incentive to manage public data, because this is less highly valued, being less personally relevant or unique (Boardman et al., 2003, Whittaker and 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 as search is often iterative involving multiple searches relating to a specific information need, some of which may result in 'deadends' (Morris et al., 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 to maintain bookmarking 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 (Wen, 2003, Morris et al., 2008). In all cases, retrieval may be made more difficult by the changing nature of the web, which may alter the content of pages users previously accessed.

We now discuss different strategies that people choose for exploitation of different types of information: namely files, emails, photos and web information.

5.2 Accessing Files

There have been significant recent developments in desktop search. 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, emails, instant messages and Web history can be accessed within the same search query. They therefore potentially 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 et al., 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 their 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 different methods (longitudinal evaluation, large scale cross sectional surveys), as well as examining 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 this self report data by collecting logfiles which allowed us to correlate self report data with actual behavioural access logs. Self reports were very accurate and highly correlated with actual behaviour, with statistical correlations being around 0.94.

We know that organisation 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 away 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 56-69% of file retrieval events and searched for only 4-16% of events. The remaining accesses were when users relied on shortcuts or used recent files to access what they had been working on. 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, despite the fact that the later version, Spotlight, was faster, as well as supporting 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 et al., (2008) asked users to characterise *exactly when* they used search as opposed to navigation and found that between 83-96% 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 for those items and their locations. And as we have seen, people are able to remember substantial amounts of information about recent files (Blanc-Brude and Scalpin, 2007, Goncalves and Jorge, 2004). The conclusion that search is used only when people can't remember the location of a file, is supported by other studies.

Jones, et al (2005) found that only 7% of users were happy with the idea that they could dispense with folders even when desktop search was available.

5.3 Accessing Email

Accessing information in email is a critical problem, given the amount of time that people spend processing it and the fact that it is both a 'todo' list for actionable information as well as an archive for more informational data (Duchenaut and Bellotti, Whittaker and Sidner, 1996, Whittaker, 2005).

A critical aspect of email management is to ensure that *actionable* items are dealt with to meet specific commitments. The previous section documented that the most common reminding strategy is to leave such items in the email inbox, hoping that these will be re-encountered on returning to the inbox to process new incoming information (Bolter, 2000, Bellotti et al., 2005, Dabbish et al., 2005, Mackay, 1988, Whittaker and Sidner, 1996, Whittaker, 2005, Whittaker et al., 2007). Variants of the 'inbox as todo list' strategy include altering the status of actionable items that have been read resetting the status of such messages so that they appear to be unread and hence bold in a standard browser (Whittaker, 2005).

Despite the central role of email in everyday work, we know relatively little about how people actually retrieve information from email. One exception is a study by Elsweiler et al. (2008) who looked at people's ability to remember emails. Participants were usually able to remember whether or not 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% of this type of information – even when items were months old. They were less good at remembering sender information, and memory for this type of information tended to decay rather quickly. Memory for temporal information was worst of all, dropping to around 50% correct over several months. In all cases, memory was affected by both the age and size of the email archive, with users remembering less when they had bigger archives or when they were required to remember older items.

Dumais et al (2003) also examined email access in Stuff I've Seen (SIS). SIS is a cross format search engine allowing users to access files, emails, web pages by issuing a query in a single interface. It also supports results sorting via attributes such as date or author. The majority of searches (74%) were focused on email 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 emails, there was a very strong focus on *recent* items, with 21% of searched for items being from the last week, and almost 50% from the last month. Many of these searches (25%) included the *name* of the email sender in the query, suggesting that (contrary to Elsweiler et al., 2008) that sender name is useful retrieval cue for emails. Elsewhere we exploit the salience of sender name in the ContactMap system which provides a specific informational view on email data, centred around network models of sender data (Whittaker et al., 2004). How can we explain the prevalence of name based search observed by Dumais et al., when compared with Elsweiler et al.'s (2008) results? Part of the difference may be due to the fact that Dumais et al. (2003) observed naturalistic behaviours which tended to be focused around retrieving recent emails. In contrast Elsweiler et al. looked at longer-term access, for more structured lab based tasks. In addition, Dumais et al. 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.

5.4 Accessing Photos

We have already described how people organise their digital pictures and the rudimentary management strategies that they employ. As with email research, there has been more focus on photo management and rather less examining exploitation. Digital photos are an extremely highly valued resource (Petrelli et al., 2008, Whittaker et al., 2010), so we should expect people to create organisations ensuring they are effective at accessing these. Indeed, work on accessing *recently taken photos* shows that people are good at retrieving these (Frohlich et al., 2002). Kirk et al (2007) asked participants to sort recent pictures in preparation for sharing these with friends or family, and found that participants were effective in finding and organising pictures taken within the last year.

These findings contrast with our own work where we looked at parents' ability to retrieve slightly older family pictures (taken more than a year ago). Despite the fact that 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 ago that they had photographed digitally. In a subsequent retrieval task, participants were asked to show the interviewer digital pictures from 3-5 of these salient past events concerning their children. To avoid having the participants choose events that they could easily retrieve, participants weren't told about the retrieval task during the initial interview. The interviewer asked participants to sit at their computer 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%). In the remainder (39%), participants simply could not find pictures of significant family events. Of the 28 unsuccessful retrieval tasks, 21 (75%) were pictures that the participants believed to be stored on their computer (or on CDs) but which they subsequently could not find. The remaining 7 were pictures participants initially thought were stored digitally, but during the retrieval process changed their minds into thinking were taken with an analog camera.

Based on participants' comments and behaviour during and after search, we identified several potential reasons for their unexpectedly poor retrieval performance: too many pictures, distributed storage, unsystematic organisation, false familiarity, and lack of maintenance. In our discussion of management we have already talked about the absence of systematic organisation 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., 2007, Rodden and 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 4475 pictures but with huge amounts of variation (SD 3039). This is a striking finding, because, consistent with other research (Kirk et al., 2005), participants all made definite efforts to reduce their overall number of pictures. For example they deleted around 17% 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 hadn't given folders meaningful names. However 67% 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 behaviours 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 organisational schemes they had imposed or software they had used.

The lack of organisation 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 name (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 accurately remember the date of the target event. Second the date label itself may be inaccurate, either because of problems with camera settings, or the folder date represents 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, moving on to other folders if they did not.

5.5 Accessing Web Documents

Accessing web pages is a problem that has been much studied. Most people's intuitions about web accesses are that these follow the pattern of *foraging*: i.e. that people predominantly seek out *new* information from the web that they then consume for the first time. The same intuitions also lead people to think that the typical way that people access web information is to rely on *search*.

One possible reason for this belief in the dominance of search is that historically web tools have 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. However, 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, and 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 analysing logfiles and history lists. Early work looking at students' browsing behaviours 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 navigate out to the various links from this page ('spokes') usually traversing no more than two links before reaccessing the hub using the 'back' button (Catledge and Pitkow, 1994). Tauscher and Greenberg (1997) instrumented browsers and looked at the rate at which people made visits to previously visited sites. They documented a recurrence rate of 58%, 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% of web actions. In addition Tauscher and Greenberg (1997) found that people were much more likely to reaccess sites that they had been to recently. Cockburn and Greenberg (2000) carried out a similar study finding a much higher frequency of accesses (81%), 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 subsequently requested them to retrieve information that they had found useful in that search session. Users were only able to successfully reaccess 20% of the sites they had visited. These users often failed to bookmark useful information believing that doing so would create 'clutter' and compromise their existing bookmark collections. Finally, consistent with other results (Teevan et al., 2004), Wen found that the general strategy for reaccess was to try and retrace prior actions, rather than attempting to search or type in prior URLs.

Aula et al (2005) looked at users self-reported strategies for web search and reaccess. They found that having multiple windows or tabs open was very common because reaccess was prevalent. In addition, the most commonly reported ways to reaccess information were to: re-access links, search for it again, directly type the URL or to save pages as local files. This confirms the results of an observational study by Bruce et al., (2004) that documented that the most prevalent strategy for refinding was to type in the URL. Other access strategies were much less prevalent, e.g. emailing links to oneself, adding URLs to a website or writing down queries. Finally there is very little use of history lists for reaccess. Aula et al. 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 et al., (2005) and Wen (2003) also noted user problems with re-access: one problem with using search to exploit information that 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 'deadends'. In trying to recover from these deadends, users often couldn't regenerate previous accesses that had been more successful. Users also couldn't recall the exact *method* that they had used for access, in consequence they had problems in 'reconstructing' search queries for information that they had originally browsed for.

In perhaps the best controlled study of revisiting, Obendorf et al, (2007) preprocessed sets of URLs for 25 users finding that revisiting rates in prior studies might have been artificially inflated by sites that automatically refreshed without user intervention. When such automatic refreshes are controlled for, they found revisitation levels were around 41%. They also documented a variety of general strategies used to access pages. The most common strategies were: using a hyperlink (44% of accesses), using forms

- including the use of search engines (15%), 'back' button (14%), opening a new tab/window (11%), typing in the URL directly (9%).

Turning specifically to revisits (as opposed to all searches), Obendorf et al. again found that the most common strategy for refinding information was to follow links (50%), with the 'back' button being the next most common strategy (around 31% of time). The remaining 'direct access' strategies (using bookmarks, homepage links, history, direct entry of URL) accounted for the final 13% of accesses. As in previous studies, re-accesses tended to be for recently accessed sites: 73% of revisits occur within an hour of the first visit, which makes the 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 aren't as reliant on 'hub and spoke' type reaccesses. They can therefore keep the context of their 'hub' page while using tabs to manage follow-up 'spoke' pages.

Finally Obendorf et al. 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% of revisits occurred within 3 mins. 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 hugely decreased, 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 link usage may reflect an orienteering strategy (Teevan et al., 2004) in which users generate plausible sets of links and then choose between 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 reaccess. 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, there are cognitive factors at work here. At medium and longer reaccess intervals users may have generated several windows or tabs, so they are unable to remember which of these they first used to access their data.

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

To summarise, then, web retrieval often involves re-accessing previously accessed data. Using links, tabs and the back button are 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 and Perez-Quinones, 2005).

5.6 Exploitation Summary

1. During exploitation, people's preference is to use manual methods (folder navigation/following links), whether this is for regular files or web data. Search is a dispreferred option even for web documents.
2. Search is not successful with personal photos (content based techniques are weak, and there is very little metadata), people therefore have to rely on browsing which turns out to be ineffective in many cases for older data.

3. Emails 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 can't 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.
4. Web Pages – despite people's intuitions, search is not the prevalent way to access web data. Reaccesses are very common with people using the back button or hyperlinks as their main reaccess methods. Reaccesses are usually for recently accessed information and the re-access strategy depends on how recently the target item was last accessed.
5. There is sometimes a mismatch 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, which means retrieval is unsuccessful for older materials. For Emails, people spend large amounts of time creating folder structures which may not always be exploited. For Web documents, people often create structures (such as bookmark collections) that aren't used because there are less costly ways to access information. They also fail to create structures that are useful.
6. Retrieval has clear regularities – there is a strong bias towards access of *recent* items, as well as a bias towards accessing a small number of items very *frequently*.

6. Future Research

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

6.1 Technology Trends

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 towards 'keeping everything' (Jones, 2004, Marshall, 2008a,b), but with interfaces that provide views onto what is important and valuable in that 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 aren't to be overwhelmed by kept data. In this spirit, we have begun to build user interfaces that keep more data (assuaging worries about deleting valuable information), 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 potential 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 organisation to generate two

tier views, and we are exploring (semi-) automatic methods for learning distinctions between these two types of information, in an attempt to reduce the burdens of manual organisation.

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 email, files and web documents. These could also include information about which information is accessed and changed most frequently. The data could then be used to generate an 'interest profile' for the user which could then be applied to incoming emails or recently accessed web pages. If for example, an incoming email or viewed webpage bears a close match to information that is already in the user's file system, then this email would be a clear candidate for keeping. In contrast, an email 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 recognising positive candidates for keeping but rather less good for deciding what should be rejected. There are various problems with automatically deleting information that is unrelated to the user's current profile. Just because incoming information is unrelated to the user's current activity, doesn't 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, and should not therefore be deleted.

Management

There is a long history of programs being built to support management (see Whittaker et al., 2007 for a review), in particular in email where many systems try to file or filter incoming emails automatically or semi-automatically. There are various problems with this approach however.

One critical problem is that users fundamentally don't trust machine learning programs (Pazzani, 2000). People are concerned that important incoming messages might be misfiled. It is clear that despite large improvements in machine learning helped by the existence of new corpora, that programs are still errorful (Whittaker et al., 2002b, Whittaker et al., 2007). And while programs promise to correctly classify documents into folders with relatively low error rates, we still lack vital empirical data about what error rates are acceptable to users. Until we clearly know whether users will at best tolerate 5% of misfiling then we don't 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 is asked to confirm or correct this. This approach is well-liked by machine learning advocates, because it provides a way to generate structured feedback on the algorithm by the user (Whittaker et al., 2004, Whittaker et al., 2007). But there is a downside to this. Unless the interface is well designed, so that suggestions and user feedback are handled in a lightweight manner, then the effort of correcting user 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 organise personal data. For example systems such as Phlat (Cutrell et al., 2006b) and Dogear (Millen et al., 2007) use social tags as ways to organise 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, and it also reduces the management costs to the individual user who has access to rich tags without generating them

him/herself. 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 analyse user activity to determine importance of, and relatedness between, documents. A common intuition is that documents that we access *frequently* are more likely to be important, as are *recently* accessed documents. The 'my recent documents' shortcut in MS Windows capitalises on the latter intuition, and more principled algorithms have also been built to capture more systematic aspects of recency (Tang et al., 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 (Kalnikaite et al., 2008a).

One specific area where machine learning might be extremely beneficial is for actionable items, which are often user's greatest concern when processing emails. Work on analysing email content has been relatively successful in predicting whether a given message requires a response (Cohen, 1996). Annotating emails with this information and presenting this in the interface might be very useful in helping people keep track of todos. Another approach to this problem is thread detection and visualisation which is now a part of newer email clients (e.g. Gmail), and research prototypes (Bellotti et al., 2003, Tang et al., 2008, Venolia and Neustaedter, 2003, Wattenberg et al., 2005) and more recent products such as Gmail. These thread viewers attempt to reduce inbox 'clutter' by clustering related messages. This has the joint benefits 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 prove to be, although one study (Tang et al., 2008) 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 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 (Kalnikaite 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. Nametags 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 number of instances of these relative strangers? And what about the success of these programs for people whose images change rapidly such as infants and young children?

Another place where machine learning has been applied is to task fragmentation. TaskTracer (Dragunov et al., 2005) is a system that analyses user behaviours in an attempt to organise them according to activities. One major problem for users is fragmentation, whereby resources relating to a common project are often placed in separate locations by applications. Thus the emails, spreadsheet, presentation and documents for a project may all be in different folders, making it hard for users to collate and organise task related materials (Bergman et al., 2004, Boardman and Sasse, 2004). Tasktracer addresses this by analysing temporal access patterns: if a webpage, document, email and spreadsheet are repeatedly open at the same time, then 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 their 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 the user's manual files, users can choose to use it if and when it offers benefits.

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 the quality of desktop search which generates too many irrelevant results. Search might be improved either by including social information (e.g. Millen et al., 2007), or more specific data about frequency and recency of document access.

There might also be different ways to view and hence access our personal information based on automatically captured data. One approach might be to project different views onto the user's data, using 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 et al., 1996a,b) 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 et al., 2002, Whittaker et al., 2004), and these systems have also proved useful as alternative email clients. And other hybrid approaches combine search with temporal landmark 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 they 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 s/he produced a document, and photos or other recordings may be available about *other activities* that the user was engaged in when that document was being worked on (Kalnikaite et al., 2008, 2010). For example, as a user might recall that they worked on a presentation for a business trip to London, and a locational view might allow them to access relevant documents by this cue. Of course there are design challenges here: there is already a huge amount of metadata available about users' activities, and interfaces will have to be carefully designed to ensure that the user is not overwhelmed by this richness.

6.2 Empirical and methodological issues

One striking observation about information curation is that we know very little about it, despite its prevalence in everyday computer use (Whittaker et al., 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 don't know much about the *relationship* between different aspects of information curation, and perhaps most importantly how management strategies influence exploitation success. What for example is the relationship between a person's folder structure and their 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 reorganise, as well as the different methods that they use to access information. At present we have only exploratory studies in these areas.

There are several practical reasons why we know so little. The first is that it is extremely hard to gather data in this area. To better understand information curation, we need to collect data about people's *personal information habits*. This is potentially intrusive, as it might require logging software to be installed on a study participant's machine, 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 needs to be *reliable* as people use it on a regular basis for everyday work. If we want users to provide feedback about a new file system, email 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 get their everyday work done (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 and 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* as 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 haven't yet transferred well to practical curation systems. For example, new algorithms are able categorise email data in standard corpora with error rates around 10%. Yet we don't know: a) what error rates users will tolerate for this type of task when carrying out everyday work; and 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 email contacts, instead preferring semi-automated methods to organise these (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).

7. SUMMARY

This review has argued that prevailing views of our information behaviours are misleading. Instead of being consumers of new public information, people's informational behaviours 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. When organising information we found surprising benefits for piles as opposed to files, although organising action-oriented information remains a major challenge. Exploitation remains reliant on manual methods such as navigation despite the emergence of desktop search. There are also mismatches between people's organisational structures and their actual retrieval requirements, e.g. for email, web documents and photos. There are a number of new technologies that could potentially address important curation problems, but implementing these in user acceptable ways remains a challenge. Finally research in this area remains in its infancy, and new data and methods are still sorely needed.

REFERENCES

Abrams, D., R. Baecker, and M. Chignell. 1998. Information archiving with bookmarks: personal web space construction and organization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 41–48. New York: ACM Press.

Ackerman, M. S. 1998. Augmenting organizational memory: A field study of Answer Garden. *ACM Transactions on Information Systems* 16(3):203–24.

Ackerman, M. S., and C. A. Halverson. 2004. Organizational memory as objects, processes, and trajectories: An examination of organizational memory in use. *Journal of Computer Supported Cooperative Work* 13(2):155–90.

Aula, A., Jhaveri, N., and Käki, M. (2005) Information search and re-access strategies of experienced web users. Proceedings of WWW 2005, May 10-14, 2005, 583-592.

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

Balter, O. 2000. Keystroke level analysis of email message organization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 105–12. New York: ACM Press.

Bälter, O., and C. L. Sidner. 2002. Bifrost inbox organizer: Giving users control over the inbox. In *Proceedings of the Second Nordic Conference on Human-Computer Interaction*, pp. 111–18. New York: ACM Press.

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

G. Bell, and J. Gemmell 2009. Total Recall: How the E- Memory Revolution Will Change Everything, Dutton.

Bellotti, V., N. Ducheneaut, M. Howard, and I. Smith. 2003. Taking email to task: The design and evaluation of a task management centered email tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 345–52. New York: ACM Press.

Bellotti, V., N. Ducheneaut, M. Howard, I. Smith, and R. Grinter. 2005. Quality vs. quantity: Email-centric task-management and its relationship with overload. *Human-Computer Interaction* 20(1–2):89–138.

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

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

Berlin, L. M., R. Jeffries, V. L. O'Day, A. Paepcke, and C. Wharton. 1993. Where did you put it? Issues in the design and use of a group memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 23–30. New York: ACM Press.

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, v.54 n.9, p.872-878,

Bergman, O., Beyth-Marom, R., Nachmias, R., Gradovitch, N., & Whittaker, S. (2008). Advanced search engines and navigation preference in personal information management. Special Issue of ACM TOIS on Keeping, Re-finding and Sharing Personal Information 26(4): pp. 1-24.

Bergman, O., Tucker, S., Beyth-Marom, R., Cutrell, E., and Whittaker, S. 2009. It's not that important: demoting personal information of low subjective importance using GrayArea. In Proceedings of the 27th international Conference on Human Factors in Computing Systems (Boston, MA, USA, April 04 - 09, 2009). CHI '09. ACM, New York, NY, 269-278.

Blanc-Brude, T. and Scapin, D. L. 2007. What do people recall about their documents?: implications for desktop search tools. In Proceedings of the 12th international Conference on intelligent User interfaces (Honolulu, Hawaii, USA, January 28 - 31, 2007). IUI '07. ACM, New York, NY, 102-11.

Boardman, R., and M. A. Sasse. 2004. "Stuff goes into the computer and doesn't come out": A cross-tool study of personal information management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 583–90. New York: ACM Press.

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

Bruce, H., W. Jones, and S. Dumais. 2004. Information behavior that keeps found things found. *Information Research* 10(1). Available at http://informationr.net/ir/10-1/paper207.html

Capra, R., and M. A. Pérez-Quiñones. 2005b. 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., W. Jones, et al., 2008. Better to Organize Personal Information by Folders Or by Tags?: The Devil Is in the Details. 68th Annual Meeting of the American Society for Information Science and Technology (ASIST 2008), Columbus, OH.

Cockburn, A. and S. Greenberg. Issues of Page Representation and Organisation in Web Browser-Revisitation Tools. Australian J. of Info. Systems, 7(2):120--127, 2000.

Cohen, W. 1996. Learning rules that classify email. In *AAAI Symposium on Machine Learning in Information Access*, pp. 18–25. Menlo Park, CA: AAAI Press.

Crawford, E., J. Kay, and E. McCreath. 2002. An intelligent interface for sorting electronic mail. In *Proceedings of the 7th International Conference on Intelligent User Interfaces*, pp. 182–83. New York: ACM Press.

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., S. Dumais, and J. Teevan. 2006a. Searching to eliminate personal information management. *Communications of the ACM* 49(1):58–64.

Cutrell, E., D. Robbins, S. Dumais, and R. Sarin. 2006b. Fast, flexible filtering with Phlat. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ed. R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson, pp. 261–70. New York: ACM Press.

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

Dragunov, A.N., Dietterich, T.G., Johnsrude, K., McLaughlin, M., Li, L., Herlocker, J.L.. TaskTracer: A Desktop Environment to Support Multi-tasking Knowledge Workers. International Conference on Intelligent User Interfaces. p. 75-82, 2005.

Drew, P. R. and M. D. Dewe. Special collection management. Library Management, 1992, 13(6): 8-14.

Donath, J. 2004. Visualizing email archives (Draft). Available from http://smg.media.mit.edu/papers/Donath/EmailArchives.draft.pdf

Dumais, S., E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins. 2003. Stuff I've Seen: A system for personal information retrieval and re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72–79. New York: ACM Press.

Ellis, D. and M. Haugan, 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 Trans. Inf. Syst. 26(4): (2008)

Farina, P. A. 2005. A comparison of two desktop search engines: Google Desktop Search (beta) vs. Windows XP Search Companion. In Proceedings of the 21st Computer Science Seminar. Hartford CT.

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

Fertig, S., E. Freeman, and D. Gelernter. 1996b. Lifestreams: An alternative to the desktop metaphor. In *Conference Companion on Human Factors in Computing Systems: Common Ground*, ed. M. J. Tauber, pp. 410–11. New York: ACM Press.

Fisher, D., Brush, A. J., Gleave E., and Smith, M. (2006). Revisiting Whittaker & Sidner's "Email Overload"; Ten Years Later, CSCW 2006.

Frohlich D, Kuchinsky A, Pering C, Don A, Ariss S (2002) Requirements for photoware, In: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work (CSCW'02), New Orleans, Louisiana, USA, New York: ACM Press, pp 166-175.
Gilbert, D. (2006). Stumbling on Happiness, Knopf.
Golder, S., Huberman, B. The Structure of Collaborative Tagging Systems, Journal of Information Science, 32(2):198-208, 2006.
Gonçalves, D., & Jorge, J.A. (2003). In An Empirical Study of Personal Document Spaces. Paper presented at the Proceedings DSV-IS'03, Funchal, Portugal.
Gonçalves, D. and Jorge, J. A. 2004. Describing documents: what can users tell us?. In Proceedings of the 9th international Conference on intelligent User interfaces (Funchal, Madeira, Portugal, January 13 - 16, 2004). IUI '04. ACM, New York, NY, 247-249.
Gwizdka, J. 2004a. Email task management styles: The cleaners and the keepers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Exmanageed Abstracts*, pp. 1235–38. New York: ACM Press.
Gwizdka, J. 2004b. *Cognitive abilities and email interaction: Impacts of interface and task*. Doctoral dissertation, University of Toronto, Toronto.
Henderson, S., & Srinivasan, A. (2009). An Empirical Analysis of Personal Digital Document Structures, HCI International 2009. San Diego, CA, USA.
Google. 2009. Google Desktop. Retrieved March 16, 2009, from http://desktop.google.com/
Hearst, M. A. 1999. User interfaces and visualization. In *Modern information retrieval*, ed. R. Baeza-Yates and B. Ribeiro-Neto. Boston, MA: Addison-Wesley.
Jones, W. 2004. Finders, keepers? The present and future perfect in support of personal information management. *First Monday* 9(3). Available at http://www.firstmonday.dk/issues/issue9_3/jones/index.html
Jones, W. 2007. Personal information management. *Annual Review of Information Science and Technology (ARIST)* 41.
Jones, W. 2007. *Keeping found things found: The study and practice of personal information management.* San Francisco, CA: Morgan Kaufmann.
Jones, W., H. Bruce, and S. Dumais. 2003. *How do people get back to information on the Web? How can they do it better?* Paper presented at the 9th IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003), Zurich, Switzerland, September.
Jones, W., and S. Dumais. 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., A. J. Phuwanartnurak, R. Gill, and H. Bruce. 2005. Don't take my folders away! Organizing personal information to get things done. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1505–08. New York: ACM Press.
Jones, W. and Teevan, J. (2007). Personal Information Management, U Washington Press.
Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision making under risk. Eugene, Ore., 1979.
Kalnikaitė, V., Sellen, A., Whittaker, S., & Kirk, D. (2010). Now Let Me See Where I Was: Understanding How Lifelogs Mediate Memory. To Appear in Proceedings of CHI 2010, ACM Press, New York.
Kalnikaitė, V. & S. Whittaker (2008a). Social Summarization: Does Social Feedback Improve Access to Speech Data? In Computer Supported Co-operative Work, ACM Press, New York.
Kalnikaitė, V. & Whittaker, S. (2008b). Cueing Digital Memory: How and Why Do Digital Notes Help Us Remember? In Proceedings of Human Computer Interaction (British HCI Conference),.
Kalnikaite, V., and Whittaker, S. (2007). Software or Wetware? Discovering When and Why People Use Digital Prosthetic Memory. In Proceedings of CHI07 Conference on Human Factors in Computing Systems, 71-80, New York: ACM Press.
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. Seattle: University of Washington Press.
Kidd, A. 1994. The marks are on the knowledge worker. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence*, ed. B. Adelson, S. Dumais, and J. Olson, pp. 186–91. New York: ACM Press.
Kobayashi, M. and Takeda, K., 2000. Information retrieval on the web. ACM Computing Surveys (ACM Press) 32 (2): 144–173
Osburn, Charles B., and Ross Atkinson. 1991. Collection Management: A New Treatise.
Greenwich, Connecticut: JAI Press.

Kirk, D., Sellen, A., Rother, C., and Wood, K. (2006). Understanding "photowork". Proceedings of CHI 2006, New York: ACM Press.

Klimt, B., and Y. Yang. 2004. *Introducing the Enron corpus*. Paper presented at the First Conference on Email and Anti-Spam (CEAS 2004), Mountain View, CA, July. Available at http://www.ceas.cc/papers-2004/168.pdf

Kobayashi, M. and Takeda, K. 2000. Information retrieval on the web. ACM Computing Surveys (ACM Press) 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. 1988a. The psychology of personal information management. *Applied Ergonomics* 19(1):55–66.

Lansdale, M. 1991. Remembering about documents: Memory for appearance, format, and location. *Ergonomics* 34(8):1161–78.

Lansdale, M., and E. Edmonds. 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. Tech. rep. Kalio, Sydney, Australia.

Lifestreams: A storage model for personal data. 1996. ACM SIGMOD *Bulletin*, March.

Lifestreams: Organizing your electronic life. 1995. AAAI Fall Symposium: AI Applications in knowledge navigation and retrieval, November, Cambridge, MA.

Mackay, W. E. 1988. More than just a communication system: Diversity in the use of electronic mail. In *Proceedings of the 1988 ACM Conference on Computer-Supported Cooperative Work*, pp. 344–53. New York: ACM Press.

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, in DLib Magazine, vol. 14, no. 3/4, Corporation for National Research Initiatives (CNRI)/ D-Lib Magazine, March 2008.

Marshall, C., 2008b. Rethinking Personal Digital Archiving, Part 2: Implications for Services, Applications, and Institutions, in D-Lib Magazine, vol. 14, no. 3/4, Corporation for National Research Initiatives (CNRI)/ D-Lib Magazine, March 2008.

Millen, D., Yeng., M., Whittaker, S., and Feinberg, J. (2007). Social Bookmarking and Exploratory Search. In European Conference on Computer Supported Co-operative Work, 179-198. Springer: Amsterdam.

Morris, D., Ringel Morris, M., and Venolia, G. 2008. SearchBar: a search-centric web history for task resumption and information re-finding. In Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 1207-1216

Nardi, B., S. Whittaker, E. Isaacs, M. Creech, J. Johnson, and J. Hainsworth. 2002. ContactMap: Integrating communication and information through visualizing personal social networks. *Communications of the Association for Computing Machinery* April, pp. 89–95.

Obendorf, H., Weinreich, H., Herder, E., and Mayer, M. 2007. Web page revisitation revisited: implications of a long-term click-stream study of browser usage. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM, New York, NY, 597-606.

Pazzani, M. J. 2000. Representation of electronic mail filtering profiles: A user study. In *Proceedings of the 5th International Conference on Intelligent Use Interfaces*, pp. 202–06. New York: ACM Press.

Petrelli, D., Whittaker, S., Brockmeier, J. (2008). Autotopography: What can Physical Mementos tell us about Digital Memories? In Proceedings of CHI08 Conference on Human Factors in Computing Systems, 53-62, New York: ACM Press.

Pirolli, P. (2007). "Information Foraging Theory: Adaptive Interaction with Information."

Pirolli, P., & Card, S. K. (1995). Information foraging in information access environments. In Proceedings of the Conference on Human Factors in Computing Systems, CHI '95 (pp. 5158). New York: Association for Computing Machinery.

Pirolli, P., & Card, S. K. (1999). Information Foraging, Psychological Review, 106, 643-675.

Ringel, M., E. Cutrell, S. T. Dumais, and E. Horvitz. 2003. Milestones in time: The value of landmarks in retrieving information from personal stores. In *INTERACT'03*, ed. G. W. M. Rauterberg, M. Menozzi, and J. Wesson, pp. 184–91. Amsterdam: IOS Press.

Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., and van Dantzich, M. 1998. Data mountain: using spatial memory for document management. In Proceedings of the 11th Annual ACM Symposium on User interface Software and Technology (San Francisco, California, United States, November 01 - 04, 1998). UIST '98. ACM, New York, NY, 153-162.

Rodden, K., and K. Wood. 2003. How do people manage their digital photographs? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 409–16. New York: ACM Press.

Rosch, E. 1978. Principles of categorization. In *Cognition and categorization*, ed. E. Rosch and B. B. Lloyd, pp. 27–48. Hillsdale, NJ: Lawrence Erlbaum.

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

Russell, D., and Lawrence, S. 2007. Search everything. In Personal information management, W. Jones and J. Teevan, eds., University of Washington Press, Seattle and London, 153-166.

Segal, R. B., and J. O. Kephart. 1999. MailCat: An intelligent assistant for organizing e-mail. In *Proceedings of the Third Annual Conference on Autonomous Agents*, ed. O. Etzioni, J. P. Müller, and J. M. Bradshaw, pp. 276–82. New York: ACM Press.

Shannon, C., and Weaver, W., 1949. A mathematical theory of communication. University of Illinois Press.

Tang, J. C., Wilcox, E., Cerruti, J. A., Badenes, H., Nusser, S., and Schoudt, J. 2008. Tag-it, snag-it, or bag-it: combining tags, threads, and folders in e-mail. In CHI '08 Exmanageed Abstracts on Human Factors in Computing Systems (Florence, Italy, April 05 - 10, 2008). CHI '08. ACM, New York, NY, 2179-2194.

Tang, J. C., Lin, J., Pierce, J., Whittaker, S., and Drews, C. 2007. Recent shortcuts: using recent interactions to support shared activities. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM, New York, NY, 1263-1272.

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, v.47 n.1, p.97-137.

Teevan, J., C. Alvarado, M. S. Ackerman, and D. R. Karger. 2004. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 415–22. New York: ACM Press.

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

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

Venolia, G., and C. Neustaedter. 2003. Understanding sequence and reply relationships within email conversations: A mixed-model visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 361–68. New York: ACM Press.

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

Wattenberg, M., S. Rohall, D. Gruen, and B. Kerr. 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., V. Bellotti, and J. Gwizdka. 2006. Email in personal information management. *Communications of the ACM* 49(1):68–73.

Whittaker, S., Bellotti, V., and Gwizdka, J. 2007. Everything Through Email. In W. Jones and J. Teevan (Eds.). Personal Information Management. Seattle: University of Washington Press.

Whittaker, S., Bergman, O., and 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., and J. Hirschberg. 2001. The character, value and management of personal paper archives. *ACM Transactions on Computer-Human Interaction* 8(2):150–70.

Whittaker, S., Hirschberg, J., Amento, B., Stark, L., Bacchiani, M., Isenhour, P., Stead, L., Zamchick G., & Rosenberg, A., 2002b SCANMail: a voicemail interface that makes speech browsable, readable and searchable. In Proceedings of CHI2002 Conference on Human Computer Interaction, NY: ACM Press, 275-282.

Whittaker, S., Jones, Q., and Terveen, L. 2002a. Contact Management: Identifying Contacts to Support Long Term Communication. In Proceedings of Conference on Computer Supported Cooperative Work, 216-225. New York: ACM Press.

Whittaker, S., Q. Jones, B. Nardi, M. Creech, L. Terveen, E. Isaacs, et al. 2004. Contactmap: Organizing communication in a social desktop. *ACM Transactions on Computer-Human Interaction (TOCHI)* 11(4):445–71.
Whittaker, S., and C. Sidner. 1996. Email overload: Exploring personal information management of email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground*, ed. M. J. Tauber, pp. 276–83. New York: ACM Press.
Whittaker, S., L. Terveen, and B. A. Nardi. 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., and Davis, M. 2004. Photo annotation on a camera phone. In CHI '04 Extended Abstracts on Human Factors in Computing Systems. CHI '04. ACM, New York, NY, 1403-1406.
Wilson, T.D., 1981. On user studies and information needs. Journal of Documentation, 37(1): p. 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. (p. 15-51) London: Aslib.
Wilson, T. 1999. Models in Information Behaviour Research. Journal of Documentation. 55(3). 249-270.