A Report[1] on the NSF-Sponsored Workshop on Personal Information Management, Seattle, WA, 2005

William Jones
The Information School
University of Washington
Seattle, Washington 98195
+1 (206) 616-1995
jones@ischool.washington.edu

Harry Bruce
The Information School
University of Washington
Seattle, Washington 98195
+1 (206) 616-0985
harryb@u.washington.edu

1 Executive Summary

A workshop on Personal Information Management (PIM) was held in Seattle, January 27-29, 2005. It was funded through National Science Foundation (NSF) grant # 0435134 [2]. More can be learned concerning the workshop (hereafter referred to as PIM2005), its structure, participants and outcomes by visiting the workshop web site: http://pim.ischool.washington.edu/

Objectives of the workshop included:

1. Examining what PIM is as a field of inquiry. What should it encompass?
2. Determining what good and better PIM looks like. How do we measure?
3. Establishing key problems and challenges that must be met if we are to make progress in PIM.
4. Identifying promising approaches to PIM (that may meet these challenges).
5. Fostering a research community for the field of PIM inquiry.

The organizational committee included: William Jones, Harry Bruce, Nicholas Belkin, Victoria Bellotti, Susan Dumais, Jonathan Grudin, Jacek Gwizdka, Alon Halevy, David Karger, David Levy, Manuel Perez-Quinones and Jef Raskin.

Personal information management or PIM is attracting increasing attention as an area of study. The payoffs for advances in PIM are large and varied.

- For each of us as individuals, better PIM means a better use of our precious resources (time, money, energy, attention) and, ultimately, a better quality to our lives.
- Within organizations, better PIM means better employee productivity and better team work in the near-term. Longer-term, PIM is key to the management and leverage of employee expertise.

Advances in PIM also translate into

- Improvements in education programs of information literacy
- Better support for our aging workforce and population.

---

[1] As a direct consequence of the PIM2005 workshop described in this report, a special section on Personal Information Management is scheduled to appear in the January 2006 issue of Communications of the Association for Computing Machinery (CACM). In addition, work is underway on a complete edited book provisionally titled “Personal Information Management: Challenges and Opportunities” with William Jones and Jamie Teevan as co-editors (through the University of Washington Press). Finally, William Jones, as sole author, is under advance contract with Morgan Kaufman/Elsevier to write a book provisionally titled, “Keeping Found Things Found: The Study and Practice of Personal Information Management”.

[2] Workshop grant was awarded through the division of Information & Intelligent Systems, in the former Information & Knowledge Management program, with Maria Zemankova as Program Officer.

Intellectual merit and broader Impacts of the workshop. Notwithstanding the obvious importance of PIM, the field of PIM research is currently fragmented. Good research relating to PIM is scattered across a number of disciplines including information retrieval, database management, information science, human-computer interaction, cognitive psychology and artificial intelligence. The workshop brought together 30 acknowledged leaders from these disciplines involved in PIM-related research as a step towards building a more integrated community of PIM research..

Special care was taken, in the selection of workshop participants and in the workshop structure to represent a diversity of user needs. For example, *personae* were selected, and used throughout workshop discussions, to represent the PIM needs and circumstances of minorities, the economically disadvantaged, students, the elderly and the disabled.

The workshop exceeded expectations as a "prime mover" in fostering a greater sense of a PIM research community. The following are among the developments that are a direct outcome of the workshop:

- A special issue on PIM in the Communications of the Association for Computing Machinery (CACM) to appear in January 2006, edited by William Jones, Jaime Teevan and Ben Bedereson. Issue editors and most authors attended the workshop. Workshop breakout session topics are covered by articles in this special issue.
- An edited book on PIM, published by the University of Washington Press, with William Jones and Jaime Teevan as editors, to appear in the Spring of 2007, provisionally titled ""Personal Information Management: Challenges and Opportunities".
- A book on PIM, published by Morgan Kaufman/Elsevier, authored by William Jones, to appear in the Spring of 2007, provisionally titled "Keeping Found Things Found: The Study and Practice of Personal Information Management".
- A special workshop on PIM to occur in conjunction with SIGIR 2006 in Seattle, WA, August, 2006.
- A full-day course on PIM with William Jones and Jacek Gwizdka as co-presenters has been accepted for inclusion in the program of CHI 2007 (http://www.chi2006.org/).

In addition, several workshop participants were featured in articles about PIM (New York Times, The Seattle Times).

The remainder of this report on the workshop is structured into sections as follows:

2. The **Introduction** provides background information on PIM as an emerging field of study.
3. **Focus Area Summaries** includes reports from breakout groups at the workshop.
4. The **Conclusion** lists some of key challenges that must be addressed if we are to make real progress in the understanding of and support for Personal Information Management.

2 Introduction

Personal Information Management (PIM) refers to both the practice and the study of the activities people perform in order to acquire, organize, maintain and retrieve information for everyday use. One ideal of PIM is that we always have the right information in the right place, in the right form, and of sufficient completeness and quality to meet our current need. Tools and technologies help us spend less time with time-consuming and error-prone actions of information management (such as filing). We then have more time to make creative, intelligent use of the information at hand in order to get things done.

This ideal is far from the reality for most people. A wide range of tools and technologies are now available for the management of personal information. But this diversity has become part of the problem leading to *information fragmentation*. A person may maintain several separate, roughly comparable but inevitably inconsistent, organizational schemes for electronic documents, paper documents, email messages and web references. The number of organizational schemes may increase if a person has several email accounts, uses separate computers for home and work, uses a PDA or a smart phone, or uses any of a bewildering number of special-purpose PIM tools.

Interest in the study of PIM has increased in recent years with the growing realization that new applications, new gadgets, for all the targeted help they provide, often do so at the expense of increasing the overall complexity of PIM. A note-taking application, for example, may provide many useful features for note-taking. But, if it also forces of a new system (e.g., tabs) for the organization of these notes that does not integrate with existing organizations for files, email messages or web references, then users can rightly complain that this is "one organization too many"

Interest in building a stronger community of PIM inquiry is further driven by an awareness that much of the research relating to the study of PIM is also fragmented by application and device in ways that parallel the fragmentation of information that many people experience. Excellent studies focus on uses of and possible improvements to email. Studies similarly focus on the use of the Web or specific web facilities such as the use of bookmarks or history information. And studies have looked at the organization and retrieval of documents in paper and electronic form.

Additional research efforts fit well under a "PIM umbrella" that maintains focus on people and what they want to or need to be able to do with their information. The completion of a task depends critically on certain information. For example, returning a phone call depends on knowing the person's first name and phone number. As such, the study of personal task management clearly relates to PIM. Research into "digital memories" and the "record everything" and "compute anywhere" possibilities enabled by advances in hardware also relate.

Good research relating to PIM is scattered across a number of disciplines including information retrieval, database management, information science, human-computer interaction, cognitive psychology and artificial intelligence. Thirty researchers from these disciplines and with a special interest in PIM met on January 27-29, 2005, at a 3-day workshop sponsored by the National Science Foundation (NSF) (http://pim.ischool.washington.edu/). A common sentiment expressed at this workshop was that research problems of PIM have often "fallen through the cracks" between existing research and development efforts.

2.1 The problems of PIM gone bad

In our real world, we do not always find the right information in time to meet our current needs. The necessary information is never found or it "arrives" too late to be useful. Information may also enter our lives too soon and then be misplaced or forgotten entirely before opportunities for its application arrive.

Information is not always in the right place: The information we need may be at home when we're at work or vice versa. It may be on the wrong computer, PDA, smart phone or other device. Information may be "here" but locked away in an application or in the wrong format so that the hassles associated with its extraction outweigh the benefits of its use. We may forget to use information even when (or sometimes because) we have taken pains to keep it somewhere in our lives. We may fail to make effective use of information even when it is directly in view.

These are failures of PIM. Some failures of PIM are memorable. Other failures may recede into a background cost of "doing business" in our world. Many of us, for example, can remember the frustration of failing to find an item of information – for example, a paper document, a digital document, an email message – that we know is "there somewhere". We may spend precious minutes, sometimes hours, in an already busy day looking for this lost information.

But even a routine day when things proceed more or less as expected is often filled with many small failures of PIM. Smaller failures may occur so often that we stop noticing them in much the same way that we may no longer notice the scuff marks on the kitchen floor or the coffee stain on a favorite shirt. These failures form a part of an "information friction" associated with our practice of PIM. A simple email request, for example, can often cascade into a time-consuming, error-prone chore as we seek to bring together, in coherent, consistent form, information that lies scattered, often in multiple versions, in various collections of paper documents, electronic documents, email messages, web references, etc.

Can you give a presentation at a meeting next month? That depends… What did you say in previous email messages? When is your son's soccer match? Better check the paper flyer with scheduled games. Does the meeting conflict with a conference coming up? Better check the conference web site

to get dates and program information. What have you already scheduled in your calendar? Can you get away with simple modifications to a previous presentation? Where is that presentation anyway? Here it is. No wait. This looks like an older version that still has some silly factual errors in it. Where is the current version?? Maybe you left it on the computer at home…

2.2 The benefits of better PIM

Information is a means to an end. Not always, not for everyone, but mostly. We manage information to be sure we have it when we need it – to complete a task, for example. Information is not even usually a very precious resource. In fact, we have too much of it. Even a document we have spent days or weeks writing is usually available in multiple locations (and, sometimes confusingly, in multiple versions). We manage information because information is the most visible, "tangible" way to manage other resources that *are* precious.

Herbert Simon elegantly expressed this point with respect to the resource of attention:

W*hat information consumes is rather obvious: it consumes the attention of its recipients. Hence, a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it. -- Herbert Simon, 1971*

The same holds true for other precious resources in our life – our time, our energy … even, our sense of wellbeing. Certainly the nagging presence of papers representing unpaid bills, unanswered letters or un-filed documents can distract, enervate and demoralize. We can't "see" our well-being or our attention or our energy or even our time (except through informational devices such as a calendar). But we can see -- and manage -- our paper documents, our e-documents, our emails messages and other forms of information.

The payoffs for advances in PIM are large and varied:

- For each of us as individuals, better PIM means a better use of our precious resources (time, money, energy, attention) and, ultimately, a better quality to our lives.
- Within organizations, better PIM means better employee productivity and better team work in the near-term. Longer-term, PIM is key to the management and leverage of employee expertise.

Advances in PIM may also translate into:

- Improvements in education programs of information literacy. Progress in PIM is made not only with new tools and technologies but also with new teachable techniques of information management.
- Better support for our aging workforce and population in order to increase the chances that our mental lifespan matches our physical lifespan.

The payoffs for better PIM may be especially large in targeted domains such as intelligence analysis or medical informatics. Better PIM may help doctors and nurses to balance a large and varied caseload. Potentially of greater importance may be PIM support for individuals undergoing long-term or sustained treatments for chronic or acute health conditions.

For example, cancer patients commonly receive a primary intervention (e.g., surgery) which is followed by subsequent therapy lasting additional weeks, months, or years. Cancer patients are frequently in the situation of managing a regimen of longer-term, outpatient care—some combination of chemotherapy, radiation therapy, hormonal therapy, additional surgical procedures—while trying to maintain their normal lives at work and at home. They are thus saddled with all normal challenges of PIM and must also manage vast amounts of new and unfamiliar information, given by range of health care professionals from a range of different organizations and departments, often only aurally, often in inconsistent forms. Moreover, patients may experience heightened, if temporary, problems with memory loss – if not a product of the treatments and operations themselves, then the product of emotional reactions (anxiety, depression) to their situations.

2.3 PIM is not new

PIM broadly defined includes the management of information going into our own memories as well the management of external information. As such, an interest in PIM-related matters is evidenced in the study of mnemonic techniques going back to ancient times.

However, although definitions of PIM vary, they generally include as a central component, the management of external forms of information. For many centuries, paper (parchment, vellum) were the primary means of rendering information in external form. As information increasingly came to be rendered in paper documents and these increased in number, so too did the challenges of managing these documents. In his autobiography, Benjamin Franklin describes his own difficulties with the attainment of the virtue of "order": "Order, too, with regard to places for things, papers, etc., I found extremely difficult to acquire".

Tools in support of the management of paper-based information were developed over time. The vertical filing cabinet that is now such a standard (if increasingly "old-fashioned") feature of offices, home and workplace, was first introduced in the early 1900s. New technologies embodied in new tools periodically spark an interest in ways of expanding the human capacity to manage and process information.

The modern dialog on PIM is generally thought to have begun with Vannevar Bush's highly inspirational article "As we may think" published as World War II was finally nearing its end. Bush recognized a problem with the sheer quantity of information being produced and with its compartmentalization by an increasing specialization of scientific discipline: "The investigator is staggered by the findings and conclusions of thousands of other workers – conclusions which he cannot find time to grasp, much less to remember, as they appear". Bush expressed a hope that technology might be used to extend our collective ability to handle information and to break down barriers impeding the productive exchange of information. Bush described a *memex* as "a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility." The memex used small head-mounted cameras to record experiences and microfilm to store these experiences, but no computer.

The phrase "Personal Information Management" was itself apparently first used in the 1980's (Landsdale, 1988) in the midst of general excitement over the potential of the personal computer to greatly enhance the human ability to process and manage information. The 1980's also saw the advent of so-called "PIM tools" that provided limited support for the management of such things as appointments and scheduling, to-do lists, phone numbers, and addresses.

2.4 A renewed interest in PIM

The past few years have seen a revival of interest in PIM[3] – not only as a "hot topic" but as a serious area of inquiry focusing the best work from a diverse set of disciplines including cognitive psychology, human-computer interaction, database management, information retrieval and library and information science.

Renewed interest in PIM is double-edged. On one side, the pace of improvements in various PIM-relevant technologies gives us reason to believe that earlier visions of PIM may actually be realized in the near future. Digital storage is cheap and plentiful. Why not keep a record of everything we have encountered? Digital storage can hold not only conventional kinds of information but also pictures, photographs, music – even films and full-motion video. Better search support can make it easy to pinpoint the information we need. The ubiquity of computing and the miniaturization computing devices can make it possible for us to take our information with us wherever we go and still stay connected to a still larger world of information. Improvements in technologies of information input and output (e.g.,

[3] For example, at CHI'2004 there were 10 full papers (out of 93), 5 short-papers, and 4 posters focused on PIM-related topics. At CHI'2005 there were 9 full papers (out of 93), 5 short-papers or posters and 1 doctoral consortium presentation focused on PIM-related topics.

better voice recognition, voice synthesis, integrated displays of information) can free us from the mouse, keyboard and monitor of a conventional computer.

This is all very exciting. But, on the other side, renewed interest in PIM is spurred by the awareness that technology and tool development, for all their promise, invariably create new problems and sometimes exacerbate old problems too. Information that was once in paper form only is now scattered in multiple versions between paper and digital copies. Digital information is further scattered into "information islands" each supported by a separate application or device. This "other side" to renewed interest in PIM recognizes that new tools, new applications – for all the targeted help they provide – can still end up further complicating a person's overall information management challenge.

3 Focus Area Summaries

The workshop included two sessions of breakouts.

- The first session (Friday morning) explored current problem areas (& opportunities). What is the current state of things? What do we know? What should we be finding out? What should things look like in the future?
- The second session (Friday afternoon) explored promising approaches to PIM. How do alternate approaches compare? What support is needed?

Specific breakout groups were as follows[4]:

Session 1. Problem areas and Opportunities.

Title	Facilitators	Participants
1. Towards a field of PIM inquiry.	William Jones, Manuel Perez-Quinones	Mike Franklin, Marcia Bates, David Levy, David Karger
2. Finding, re-finding, reminding and "re-collection" of personal information.	Jaime Teevan, Nick Belkin	Rick Boardman, Ofer Bergman, Jacek Gwizdka, Ben Bederson
3. Encountering, keeping, organizing & maintaining information	Cathy Marshall, Harry Bruce	Brian Ross, Tiziana Catarci, Doug Gage, David Maier
4. From PIM to "GIM".	Tom Erickson, Jonathan Grudin	Steve Whittaker, Sue Dumais, Alon Halevy
5. Measurement and evaluation.	Diane Kelly	Wanda Pratt, Jim Gemmell, Mary Czerwinski

Session 2. Promising approaches.

Title	Facilitators	Participants

[4] Several additional topics were discussed in the large group meetings of Thursday afternoon and Saturday morning including:
11. Special groups, special problems and "deep" applications of PIM including patient PIM
12. Teachable strategies of PIM
13. The uses of a database structure in PIM
14. The use of blogs and wikis in PIM.
15. Uses of semantic web initiatives, in particular developments XML and RDF, in PIM
16. The evaluation of PIM tools
17. The role of schema and classification schemes in PIM

6. Towards a unification & integration of PIM support.	David Karger, William Jones	Ofer Bergman, Wanda Pratt, Mike Franklin
7. Enhancements of personal information.	David Maier, Alon Halevy	Marcia Bates, Ben Bederson, Harry Bruce
8. Search, finding, filtering and auto-classification.	Nick Belkin, Susan Dumais, Diane Kelly	Jaime Teevan, Rick Boardman, Brian Ross
9. Digital memories, ubiquitous computing	Mary Czerwinski, Jim Gemmell, Doug Gage	Cathy Marshall, Tiziana Catarci, Manuel Perez
10. Beyond email…	Steve Whittaker, Jacek Gwizdka	Tom Erickson, Jonathan Grudin, David Levy

3.1 Towards a field of PIM inquiry

William Jones, Manuel Perez-Quinones, Marcia Bates, Mike Franklin, David Karger. David Levy, Mel Knox (student volunteer).

The discussion group was organized to consider key questions relating to PIM as a field of inquiry including:

1. What does it mean for PIM to be a field of inquiry? What does it take? Is this necessarily what we want?
2. What is PIM (at its core) and isn't? What are its components?
3. Is there a conceptual framework which might help (as a way to approach PIM and its components)?
4. How do we measure progress in PIM as a field? With what benchmarks?

Since another discussion group led by Diane Kelly was also consider question #4, our group focused on the first 3 questions. We considered the possibility that what's needed for progress in PIM is a community – from this may emerge a field over time. There is no field. A community is emerging. There is certainly interest and a need. The group agreed that many important PIM concerns were currently "falling through the cracks" . On the other hand, PIM as an area of study, provides a good meeting ground and area of application for the work of several different disciplines including information retrieval, database management, artificial intelligence, human-computer interaction and cognitive science.

The discussion then moved to a consideration of what PIM should encompass as an area of study. PIM is a large area with uncertain boundaries. It includes all efforts to work with, deal with, and react to information at a personal, individual level. PIM includes various activities to search for, find, encounter, interpret, decide to keep (or not), file and organize for re-use, re-access and ultimately use information. Good, timely information is critical to a wide range of tasks, professional and personal.

A deeper understanding of what PIM is, at its core, and at its broad periphery of overlap with other fields of inquiry, begins with consideration of definitions for PIM and associated concepts.

3.1.1 Some working definitions

Definitions offered here are "working" in their intended primary purpose to further the article's exposition. It is recognized that alternate, often better, definitions can be formulated for each concept and it is quite beyond the scope of this report to consider these alternatives.

3.1.1.1 Information and the information item

The statement above holds in particular for "information". In this report on PIM, we focus especially on the capacity of information to affect change in our lives and in the lives of others. The information we

receive influences the actions we take and the choices we make. We decide, for example, which of several hotels to book depending upon the information we are able to gather concerning price, location, availability, etc. Incoming information helps us to monitor the state of our world. Did the hotel send a confirmation? What about directions?

We also send information to affect change. We send information in the clothes we choose to wear, the car we choose to drive, and in the way we choose to act. We send information (often more than we intend) with every sentence we speak or write. It is with respect to the information we send, that it is most clearly necessary to go beyond Shannon's original notions of information as a collaborative exchange between sender and recipient. As Machiavelli might have said, we send information to serve our own purposes. Certainly one of these purposes is to be helpful and inform others. But we also send information to persuade, convince, impress and, sometimes, to deceive.

An ***information item*** is a packaging of information. Examples of information items include: 1. paper documents. 2. electronic documents and other files. 3. email messages. 4. web pages or 5. references (e.g., shortcuts, alias) to any of the above. Some might prefer to use the term "information object" to emphasize the point that an information item can be acted upon. Items encapsulate information in a persistent form that can be created, stored, moved, given a name and other properties, copied, distributed, deleted., moved, transformed, etc.

The support that we depend upon for our interaction with paper-based information items includes our desktop, paper clips, staplers, filing cabinets, etc. In our interactions with digital information items we depend upon the support of various computer-based tools and applications such as an email client, the file manager, a web browser, etc. The "size" of current information items is partly determined by these applications. There are certainly situations in which some of us might like an information item to come in smaller units. A writer, for example, might like to treat paragraphs or even individual sentences as information items (to facilitate their re-use). A salesperson might view the individual entry in a contact management database as an information item. Applications exist in each case to help (e.g., contact management software, writer's software such as DevonThink)..

An information item has an associated ***information form*** determined by the tools and applications that are used to name, move, copy, delete or otherwise organize or assign properties to an item. The most common forms we consider in this report are paper documents, e-documents and other files, email messages and web bookmarks.

It is striking to consider how much of our interaction with the world around us is now mediated by information items. We consult the newspaper or, increasingly, a web page to read the headlines of the day and to find out what the weather will be like (perhaps before we even bother to look outside). We learn of meetings via email messages. We receive the documents we are supposed to read for this meeting via email as well.

On the sending side, we fill out web-based forms. We send email messages. We create and send out reports in paper and digital form. We create personal and professional web sites. These and other information items serve, in a real sense, as a proxy for ourselves. We project ourselves and our desires across time and space in ways that would never have occurred to our forbearers.

Another point concerning information items, in contrast, for example, to what we hear or see in our physical world, is that we can often defer processing until later. We can, and do, accumulate large numbers of information items for a "rainy day". This is quite unlike, for example, the scenarios of situation awareness where acceptable delays in processing information are measured in seconds.

Finally, there is sometimes discussion of Personal Knowledge Management (PKM). Given the usual ordering of data < information < knowledge, we are tempted to think that PKM is more important than PIM. That may be so. One major challenge of PKM, just as with knowledge management more generally, is in the articulation of rules and "lessons of a lifetime" in a form that we (and possibly others) can understand. Knowledge expressed and written down becomes one or more items of information – to be managed like other information items.

3.1.1.2 Personal information.

The discussion group considered several senses of p*ersonal information*:

1. The information people keep for their own personal use.
2. Information about a person but possibly kept by and under the control of others. Doctors and health maintenance organizations, for example, maintain health information about us.
3. Information experienced by a person even if this information remains outside a person's control. The book a person browses (but puts back) in traditional library or the pages a person views on the Web are examples of this kind of personal(ly experienced) information.

This report is (like the workshop) primarily concerned with the first sense of "personal information". However, we consider the 2nd sense of "personal information" in the context of an all-too-brief discussion of privacy and security. We consider the 3rd sense of personal information briefly as part of a later discussion of effort to personalize a person's experience of the web and web search.

The third case -- information we experience but do not keep in our PSI – can sometimes pose a special kind of PIM problem: We remember the information, but maybe not enough about the information to be able to find it again later. For example, we might see information on a web site about a concert in another city by our favorite musical group. But since we can't attend, we take no special steps to keep this information. Later, we find out that we must attend a business meeting in that city on the same week as the conference. We want to get back to the web page but can't recall how we got there to begin with and can seem to formulate a query to return the web page as one of the results (we're not sure since we don't know what web site's name is or how it would appear in the listing of results).

3.1.1.3 A Personal Space of Information

A personal space of information (PSI) for a person includes all the information items that are, at least nominally, under that person's control (but not necessarily exclusively so). A PSI contains a person's books and paper documents, email messages (on various accounts), e-documents and other files (on various computers). A PSI can contain references to web pages. A PSI also includes applications, tools (such as a desktop search facility) and constructs (e.g., associated properties, folders, "piles" in various forms) that support the acquisition, storage, retrieval and use of the information in a PSI.

A few other things to note about a PSI:

- Although we have some sense of control over the items in a PSI, this is partly illusory. For example, an email message can be deleted so that it no longer appears in an inbox. However, the message is very likely still around somewhere (as some have learned to their chagrin).
- A PSI does not include the web pages we have visited but may include copies (in a cache) and does include the bookmarks we create to reference these pages.
- Does PSI include our own internal memories? On the one hand, the answer must surely be yes. What could be more personal? No one else owns our memories but us. But, paradoxically, an argument can be made for "no". How much control do we have over what goes into our memories? Or what comes back out? Some things lodge in our minds even though we wish they would not. We cannot forget, i.e. we cannot simply press a "delete" key.
- In general, there are large unavoidable grey areas. For example, the files we place on a network share should probably be considered a part of our PSI even though they may not be under our exclusive control. Similarly, a PSI should probably include the many icons that applications like to leave on our computer desktops and the bookmarks and folders that are automatically created.
- A PSI is, by definition, "everything". We each have only one PSI.
- A PSI is distinguished from a Personal Information Environment (PIE) which, as used in the literature, commonly refers to subset of a PSI in combination with supporting tools. The physical space of an office including papers piled and filed, the stapler, filing cabinets, etc. is a PIE. A notebook computer is a PIE. A person can have several PIEs.

- The size of our PSI continues to grow, especially with respect to digitally encoded information. The PSI is a potential source of information for use a number of different ways. The PSI might be used, for example, to customize our experience of the Web (see the section below on finding/re-finding). The information of a PSI might be "mined" to extract important patterns in our information (and our interactions with this information). Effective re-use of the information in the PSI promises to improve our productivity. At the same time, the growing size of our PSI also raises serious questions of privacy and security.

3.1.1.4 Definitions of Personal Information Management

PIM is easy to describe and discuss. We all do it. We all have first-hand experiences with the challenges of PIM. But PIM is much harder to define. PIM is especially hard to define in ways that preserve focus on essential challenges of PIM.

Lansdale (1988) refers to PIM as "the methods and procedures by which we handle, categorize, and retrieve information on a day-to-day basis". Barreau (1995) describes PIM as a "system developed by or created for an individual for personal use in a work environment". Such a system includes "a person's methods and rules for acquiring the information …, the mechanisms for organizing and storing the information, the rules and procedures for maintaining the system, the mechanisms for retrieval, and procedures for producing various outputs".

Boardman (2004) notes that "Many definitions of PIM draw from a traditional information management perspective – that information is stored so that it can be retrieved at a later date".

In keeping with this observation, as exemplified by Barreau's definition, we might analyze PIM with respect to our interactions with a large and amorphous PSI. From the perspective of such a store, the essential operations are input, storage (including organization) and output.

In rough equivalence to input-storage-output breakdown of actions associated with a PSI, the group considered a conceptual framework with the following grouping of essential PIM activities:

- **Keeping** activities affect the input of information into a PSI.
- **Finding/re-finding** activities affect the output of information from a PSI.
- **"M-level activities"** (e.g., "m" for "mapping" or for "maintenance and organization") affect the storage of information within the PSI.

This framework was discussed in the group and then presented to the larger group of all workshop participants. There was consensus to elaborate upon this framework for the final report. The following section makes a first attempt to do this.

3.1.2 A conceptual framework and focus: PIM activities that map between information and need

Note: the following section is a substantial elaboration on the keeping/finding/m-level framework initially discussed in the discussion group and then presented to the entire

The remainder of this report's content and organization are guided by a conceptual framework that derives from a basic assumption concerning PIM activities:

PIM activities are an effort to establish, use and maintain a mapping between information and need.

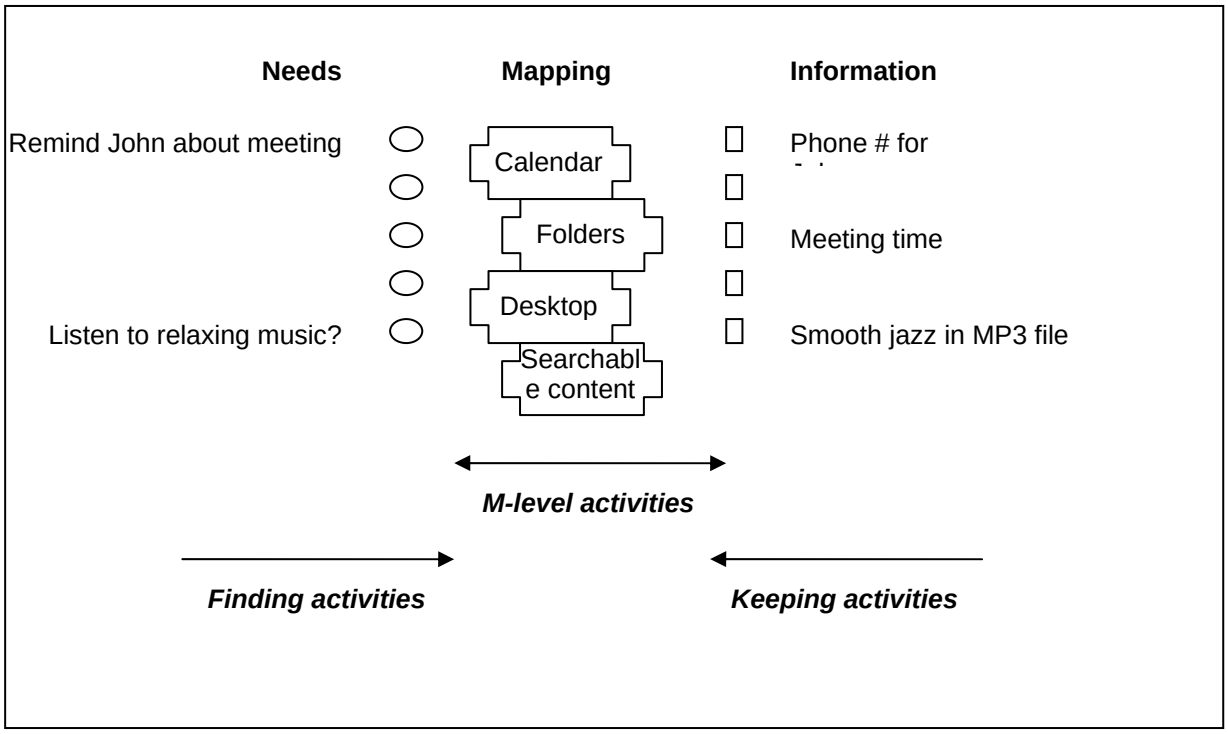

Figure 1. PIM activities viewed as effort to establish, use and maintain a mapping between needs and information

This simple statement can be expanded with reference to the diagram of Figure 1. Examples of information as listed in the rightmost column are expressed in various ways – as aural comments from a friend or colleague, as a billboard we see on the way to work or a message we hear over the radio and via any number of information items including documents, email messages, web pages and, even, hand-written notes.

Needs, too, as depicted in the leftmost column, can be expressed in several different ways: The need may, more or less, come from within us as we recall, for example, that we need to make plane reservations for an upcoming trip, or it may come via the question of a colleague in the hallway or a boss's request. Needs are often themselves packaged in information items such as email messages and web-based forms.

Connecting between need and information is a mapping. Only small portions of this mapping have an observable external representation. Large portions of the mapping are internal to our own memories – memories for specific experiences with information, experiences with information sources and kinds of information and, more broadly, our memories for the fabric of the world around us, its conventions, its "language" – it all goes into the mapping. Large portions of the mapping are potential and not realized in any form – external or internal. A sort function or a search facility, for example, has the potential to guide us from a need to desired information.

But parts of the mapping can be observed and manipulated. The folders of a filing system, digital or paper-based, the layout of a desktop, physical or virtual, the choice of file names and other properties for information items – each forms a part of an observable fabric helping to knit need to information.

PIM activities can be grouped, with reference to Figure 1, according to whether the initial focus is on a need, information or the mapping between need and information:

3.1.2.1 From need to information

find(need) -> information

We have a need. We try to find information to meet that need. Needs can be large and multi-faceted – the need for information for a review, for example – or small and simple – a locating someone's phone number. Needs can originate, more or less, in our own heads or they can come from outside – a hallway request from a colleague, for example. Frequently, a need, itself, comes packaged in an information item – an email request, for example, or via a web-based form requesting certain information for its completion. A need frequently equates with or is a part of a task (e.g., "prepare for the meeting", "answer my boss's email", "return the client's call"). But other needs may not fit tasks except by the broadest definition (for example, "see that funny web site again" or "hear 'Five to one' for old time's sake").

In our efforts to meet a need, we seek. We search. We browse. We scan through a results list or the listing of a folder's contents in an effort to recognize information items that relate to a need. Especially important, we remember to look in the first place. Sometimes, the information comes from our PSI and is information we've used many times before. Other times, the information comes from the Web and is new to us.

These activities are all referred to in this report as ***finding*** activities. "Finding" places the emphasis on the outcome – information meeting the need is "found" rather than the process. "Finding" includes "re-finding". We may repeat many of the same steps in an act of re-finding that we took to find the information in the first place. If a web search worked the first time, we may use much the same search with the same search terms a second time, for example. Finding is meant to include various information seeking activities.

Finding also applies, of course, to activities that target physical objects in our world and, as such, invites some interesting comparisons between our physical and digital worlds. We try to find a can opener that is, we think, somewhere in our kitchen. Or we may look in our closet for a pair of shoes that we want to wear to a dinner party.

Information items may occupy a virtual space, but such a space cannot, yet, compete with the richness of our physical spaces. On the other hand, we can search for digital information using computer-based tools in ways that we cannot (yet) use for the search of physical objects. But there are many similarities as well. We can fail to find an ingredient – walnuts, for example – that might be perfect for a salad we're making for any of several reasons – each with their digital analog. The walnuts may not be on the shelves we look through in the kitchen. Perhaps they aren't in the kitchen at all. Or the walnuts may be right there in front of us on the shelf but in a container that we do not recognize. Or, in the midst of everything else we are doing to prepare dinner, perhaps we forget to look for the walnuts – this, too, is a failure of finding. Similarly, we may fail to find a web site we have bookmarked for our current project for any of several reasons. We may look in the wrong folders, or perhaps the bookmark is on another computer entirely, or we may fail to recognize the bookmark though is there in front of us. Or, especially in our rush to complete the project, we may forget about entirely about the bookmark.

Finding is broadly defined to include both acts of new finding where there is no previous memory of having the needed information and to include acts of re-finding. The information found can come from inside or outside a PSI. More broadly still, finding includes efforts to create information "from scratch" as in "finding the right words" or "finding the right ideas". When crafting an information item – a "simple" email response or a much longer, more structured document – we have many choices. We have choices concerning what information is referenced and from where. For example, is it faster to look for a bookmark in our PSI that points to needed information on the Web or is it faster to simply search again using our favorite web search service? We have choices concerning how much of the item is "old" – composed with reference to, and perhaps a copy and paste from, other documents we have previously authored – and how much is "new" – coming directly from our own minds and through the keyboard without (conscious) reference to previous information. Our choices reflect often complicated calculus of expected cost and expected benefit.

Several questions arise concerning the actions of finding vs. re-finding how these might change when the target is information inside a PSI vs. information on the Web or elsewhere "out there" (see also the breakout report on finding, http://pim.ischool.washington.edu/breakouts.htm). Several studies indicate an enduring preference for browsing (e.g., by going through a nesting of folders) as a means of return to information within a PSI. Certainly, search is widely used on to locate information on the Web but,

search is often used in common with a hyperlink-enabled kind of browsing. For example, a we might use a search service to find a web site of interest and then browse within the site to locate web pages containing specific information of interest. Some members felt that there wasn't much difference between finding activities targeting new information on the Web vs. re-finding activities directed towards information in the PSI that has already been experienced. However, other breakout participants noted that there is often a strong emotional component associated to attempts to re-find information in our PSI. We may get frustrated, feel back or feel that we're "losing control" when we can find information that we "know is in there somewhere". For items repeatedly re-accessed and re-used there is also a question concerning when we "keep" this information in a way that makes a subsequent effort to re-find (re-access) easier. For example, if we go the same web site via Google search two or three times a week and then follow hyperlinks to a specific page at that site, when do we decide to "keep" the page in some other way – by making a link or a bookmark, for example – so that re-access is faster?

3.1.2.2 From information to need

keep(information) -> need

Many events of daily life are roughly the converse of finding events: Instead of having a need for which we seek information, we encounter information and try to determine what, if anything, we need to do with this information. We encounter information in many different ways and forms. We come across an interesting announcement for an upcoming event in the morning newspaper. A colleague at work may whisper news of an impending re-organization. An email may arrive with an announcement or a "for your information". While searching or "surfing" the Web for one need, we frequently encounter information that might be useful for some future need.

Decisions and actions relating to encountered information are referred to in this report collectively as ***keeping*** activities. Is the information at all relevant or potentially useful? Do we have an anticipated need for this information? We can safely ignore much of the information we encounter – the likelihood that we will need it is small and the cost of not having the information is small as well. Other information can be "consumed" immediately with no need to make special efforts to connect this information to need. Sport scores, weather reports, and stock market reports fall into this category.

There is then a middle area of encountered information. We may have a need for this information, but not now. We must then decide whether to keep this information and, if so, how. Even if we judge the information to be useful, we may still decide that no special action is required – perhaps because we already "have" this information somewhere in our PSI or because we can easily return to the information, for example, by repeating the same search or the same path of hyperlinks that brought us to the information in the first place.

If we decide to keep the information we have encountered, then we must decide how. Information kept wrong may be useless when a need for it arises later on. In worst case, we may forget about the information entirely.

As an example, a salesperson gives us her business card that includes her phone number. Do we need to keep this information at all? The answer may be "no", either because we don't care to contact this person again or because we're certain we can easily access her phone number by another means – web lookup or via a friend or colleague, for example. On the other hand, we may decide this information is important enough to keep in several different ways. We may write the phone number down in a notebook or in a calendar to be sure of calling her again later. We may also enter this information into a contact database. But none of these methods of keeping may be any good to us if we're stuck in traffic and want to call her on our mobile phone to tell her we're running late to the meeting. (If only we had also entered the number into our phone…).

Keeping activities must address the multi-faceted nature of an anticipated need. When and where will we need the information? We must also assess our own habits and anticipate our own state of mind. Will we remember to look? Will we remember to look in this particular folder? Will we recognize the information? Will we even remember why we kept it?

If our information is fragmented between devices and applications we must also anticipate the form in which we will need the information. On which device? (Laptop or mobile phone?) In which application?

As the example of the phone number illustrates, the number of ways to keep information has grown considerably in recent years as part of an overall increase in the number of devices and applications that we depend upon to manage our information. Paper is still very much a part of people's lives. In addition, we now manage electronic documents and other computer-based files, web references (as bookmarks, for example) and, of course, large numbers of email messages often in multiple accounts. We have desktop computers, laptop computers, smart phones, PDAs and ordinary notebooks.

There are many variations in keeping. We "keep appointments" by entering a reminder into a calendar. We keep good ideas that occur to us or "things to pick up at the grocery store" by writing them down in a notebook or on a loose piece of paper. We frequently re-keep information inside our PSI. For example, as we encounter a forgotten web bookmark during a "spring cleaning", we may decide to move the bookmark to a new folder where we are more likely to notice it in the future. Or, as we comb through the documents associated with a completed project, we may decide that some of these documents still have value in connection with a new project and should either be moved to a corresponding folder or assigned a label for this new project.

3.1.2.3 A focus on the mapping between need and information

A third set of PIM activities is focused on the mapping that connects need to information. These are collectively referred to in this report as ***m-level*** activities. "M" as in "mapping" or "meta". "M" also as in "maintaining and organizing", "managing" (access to and distribution of the information in PSI), "measuring" (the effectiveness of a mapping and the structures, strategies and supporting tools associated with its creation, use and upkeep). And, possibly, also "M" as in "manipulating and making sense" of a PSI and its information. Each of these senses of "M" is now described in more detail.

- ***Mapping.*** As noted earlier, only small portions of the mapping for a PSI have an observable external representation. Large portions of the mapping are internal to our own memories – memories for specific experiences with information, experiences with information sources and kinds of information and, more broadly, our memories for the fabric of the world around us, its conventions, its "language" – it all goes into the mapping. Large portions of the mapping are potential and not realized in any form – external or internal. A sort function or a search facility, for example, has the potential to guide us from a need to desired information.
- ***Meta.*** One m-level activity is to "step back" and think about the mapping overall or for a subset of the PSI (e.g., the files on a laptop computer). How should information be structured? According to what schema? For common forms of information, this means deciding on a folder structure. But in the future we may also be able to organize items according to a rich set of properties. Certainly a challenge in such a property-based system will be to select properties that truly distinguish among the items, current and likely, without creating a lot of extra work. It is at the meta level that we also consider the potential utility of supporting tools that are proffered to help us. And we also consider strategies of PIM ("file everything right away", "don't file anything", "keep everything", "don't' keep any paper", etc.)

By analogy, in we may think of or read about a great new way to organize our kitchen or our clothes closet. We may even consider a re-model that gives us more space or the purchase of a "tool" (e.g., a "drawer-design" refrigerator or stacking boxes for our clothes).

- ***Maintaining and organizing.*** We implement our "meta-level" scheme of organization through the actual creation of folders and a folder hierarchy (or through the creation of properties). Periodically, this structure needs to be updated. Some folders, for example, may no longer be needed. Some folders have grown too large and may need to be divided into subfolders. Folders may need to be moved or re-named. Information items themselves may similarly need to be deleted or moved.

By analogy, the food and utensils of a kitchen or the clothing of a closet may occasionally need to be re-distributed. We may also periodically attempt to weed out older items for donation that are no longer in use.

Maintenance of a PSI also includes updating of information content as well as organization. When much the same information is scattered in different forms and many variations through a PSI, updating can be extremely difficult. If, for example, a friend's email address or phone number or, worse, name

changes, older information items with the incorrect information (e.g., birthday reminders, directions, holiday card lists, etc.) may linger for months or years after our initial efforts to update. More extreme, are situations for large parts of our personal information are rendered irrelevant or "wrong" by a new event. People recovering from a heart attach, for example, may want a radical update in their PSI to reflect the radical change they hope to accomplish in their style of life.

- ***Managing*** *privacy, security and the distribution of items in a PSI.* A discussion of privacy and security brings us back again to a consideration not only of "our information" but also information "about us" and the large overlap between these two kinds of personal information. If our first reaction is to say "personal information is personal and no one else can see it" we are likely to have a later realization that some distribution of our personal information can be very useful. We want the travel agent to know about our seating preferences. We want colleagues and friends to know about our schedule. We may want close friends and family to know about our current condition if we are battling a serious illness. The increasing use of the personal web sites as a means to publish (and project) naturally brings a desire for technology that can support a "personal policy on privacy and security" that allows for finer distinctions that "everyone can access" or "no one can access". But, given this greater control, there is a need for user interfaces that can guide us in our choices and make clear their implications.
- ***Measuring*** *the effectiveness of a mapping and the structures, strategies and supporting tools associated with its creation, use and upkeep.* We must periodically ask ourselves "is it working?" Are the structures we've selected maintainable? Are the strategies we try to follow sustainable? Is this tool really helping or is it more trouble than it's worth? For paper documents, the signs that "things aren't working" are sometimes all too clear. For example, if paper documents continue to pile up in a "to be filed" stack and we never have time to actually file these documents away, this may be a sign that our "great new organization", for all its promise, is simply not sustainable. The signs for digital information may be more subtle. As we look for efficient, accurate, objective ways to evaluate our own practice of PIM we run into many of the same problems, at an individual level, that are also in evidence for the field of PIM. We return to this topic in the next section and also in a later section on the methodologies of PIM.
- ***Manipulating and making sense*** *of our information.* As we consider a collection of information, what are we seeing? What do we have? The folders that we still use as perhaps the most common way of organizing information items can also obscure. They can create barriers within a PSI not unlike the barriers Bush observed between an ever increasing number of scientific specializations: "*publication has been extended far beyond our present ability to make real use of the record. The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships." (Bush, 1945).* The wording in these sentences needs only slight modification to apply equally to the prodigious amounts of information we are able to store in a PSI. And we might indeed complain that the tools we have available for manipulating and making sense of, for example, a collection of computer-based files has changed little over the past two decades.
- ***Mañana****? Or maybe tomorrow (but not today).* We might also say, jokingly but with considerable truth, that "m" stands for "maybe tomorrow but not today". The m-level activities described here are easy to avoid and put off. None of them demand our attention in the way that an immediate need or even encountered information do. We perform activities of finding and keeping throughout a typical day. M-level activities can and are postponed for weeks on end. And then there is that messy closet…. Part of the problem is that we prefer to pay the incremental, perhaps barely noticeable, costs associated with the use of a poor mapping rather to suffer the certain and immediate costs of an m-level activity.

References

Barreau, D. K. (1995). Context as a factor in personal information management systems. *Journal of the American Society for Information Science, 46*(5), 327-339.

Boardman, R. (2004). *Improving Tool Support for Personal Information Management.* Imperial College, London.

Bush, V. (1945, July 1945). As We May Think. *The Atlantic Monthly*.

3.2 Finding, re-finding, reminding and "re-collection" of personal information.

Jaime Teevan, Nick Belkin

Rick Boardman, Ofer Bergman, Jacek Gwizdka, Ben Bederson

Don't Drop the Ball: Re-finding Personal Information

Personal information management has been described as a game of catch, where a person tosses their personal information into the future, in hopes of being able to catch the information later when it is needed [2]. This report focuses on the catching aspect of personal information management, discussing current approaches to and problems with how people return to previously encountered information when it has become useful.

As an example, imagine Alex, who organized his company's football team several years ago, and was recently asked by the current captain where he purchased the team jerseys. Alex must use his memory and the organizational structure he created when managing the football team to re-find the company's name. He could use the structure he created to help him remember the company name by searching, for example, for email communications with the jersey producing company in an old email folder or for the invoice in a file directory. He could also search the Web, using an old bookmark to return to the company's Web page or issuing a search to an Internet search engine for something like, "football jerseys" and browsing the result list for a familiar looking Web page.

However, the organizational structure Alex created several years ago is probably difficult for him to operate in effectively now. Further, the bookmarks Alex made are likely to have changed, and he is unlikely to be able to recognize the company's Web page should it be presented to him, let alone be able to issue an Internet search that finds it. His hunt for the name of the company where he purchased the team jerseys will probably require significant effort, and if he is unable to find the name, the current captain will be required to also expend significant effort repeating research that Alex has already performed.

As can be seen from this example, re-finding personal information is an important problem that is difficult to solve. The amount and types of information that people routinely encounter, create, use and/or save in digital form are expanding dramatically. We can assume that this increase will continue, as computing becomes ever more ubiquitous and part of our daily lives, creating a great need for effective re-finding solutions. Current tools for re-finding even textual personal information are only in their infancy, and are based on rather traditional information retrieval models, without taking into account the particular characteristics of the personal information situation [4].

Below we discuss several important controversial statements on the topic of re-finding, highlighting key arguments for and against each statement. Short term and long term goals that arise from the statements are highlighted, as are any resources needed to pursue resolution of the controversy.

Note that in this report, terms such as *search* or *finding* do not refer exclusively to keyword search (*e.g.*, Alex's Internet search for "football jerseys"), or even directed search (*e.g.*, Alex's search for the company name), but can also refer to the entire information seeking process (*e.g.*, the new captain's effort to learn about a good company from which to purchase new football jerseys) [1].

Finding = Re-finding

2 participants agree, 4 disagree

The first controversial statement is that re-finding is essentially the same behavior as finding. In this section we discuss whether we believe the two behaviors are the same, and, if they differ, what the important aspects of that difference are. A significant feature of re-finding is that people tend to know a lot of meta-information about the item they are seeking. For example, if Brooke wanted to purchase a CD she saw earlier on Amazon.com, she would probably return to Amazon.com, and use information about how she originally encountered the CD to follow a similar path to return to it.

Nonetheless, the strategies that people tend to employ when searching for new information versus returning to previously viewed information appear to be similar. Teevan, et al. [6], found that regardless of whether people were looking for information on the Web (usually a finding behavior) or in their files and email (usually a re-finding behavior), they tended to navigate to their information target via a series of small steps, using the various meta-information they knew about that target to inform the steps. For example, even if Brooke were searching for a new CD to purchase, she might know the basic genre of music she likes, what sort of CD cover art tends to appeal to her, and that Amazon.com is a store where music CDs can be purchased. She could use this information to find the CD by visiting Amazon.com, navigating to her preferred music genre, and then browsing for appealing cover art.

This finding behavior is very similar to the previous example where Brooke was re-finding, except that when finding for the first time she does not have personal experience with the actual target. Instead, the meta-data she uses is based on prior experience with similar items. Those who argue there is a difference between finding and re-finding claim that there is a qualitative difference between the meta-data a user knows about their information target based on experience with the actual item, such as exactly what the item looked like or when it was last seen, and other types of meta-data a user might have about an unknown target.

It has also been argued that when searching, a person experiences considerably more frustration when unable to locate the target if the target has been seen before than if it has not. For example, Brooke is likely to find it more frustrating to not be able to return to a CD she's already seen on Amazon than to not be able to find a new CD she likes. The amount of frustration a user experiences with a search is probably related to the searcher's expectation that the item exists, but whether a person can or cannot have a similar level of expectation for an item that has not been seen before as for an item that has is a matter of debate.

Another difference cited between finding and re-finding is that users can easily recognize their target when it has been seen before, rather than having to think about and determine that a particular item is indeed what is being sought. Those who believe re-finding and finding are the same again believe that sometimes items that have been seen before can take effort to recognize, while new items might be immediately recognizable as relevant.

Those participants that believe finding and re-finding are the same, believe that the same tools should support both behaviors. However, if re-finding is indeed found to be qualitatively different, it remains an open question as to whether the two behaviors should be supported differently, and, if so, how.

Re-finding in Personal Information = Re-finding on the Web

6 participants agree, 0 disagree

Regardless of whether finding and re-finding are the same behavior, we also discussed whether there was a qualitative difference in re-finding behavior based on the corpus. Although there is consensus among the participants in our breakout group that re-finding in one's personal information space is the same as re-finding on the Web, this statement is not taken to be true among the general PIM community. It is our belief that once information has been seen, it enters a person's personal information space, regardless of whether that information continues to reside on the Web or under direct control of the individual.

This is an open question about PIM in general, and not necessarily unique to the problem of finding and re-finding information. Many believe that there is a fundamental difference between information one believes they have control over and information that others have control over, and while this is likely to be true, the degree to which this makes *re-finding* qualitatively different in the two situations is unclear to us.

People shouldn't have to do any work in advance to make re-finding easier.

5 participants agree, 1 disagrees.

Earlier, personal information management was described as a game of catch. But should it really be necessary to "toss" information into the future in order to be able to catch it when needed? Or should relevant information be provided to a user regardless of their previous interaction with that information?

As an example, Alex, mentioned above, might have created considerable organizational structure when organizing his company's football team. This structure would serve useful to him when later asked to re-find the name of the company he ordered jerseys from. If he does not have a rich organizational structure, he might have a more difficult time re-finding that information. Organizational structure allows the user to use recognition, rather than recall, in their search process.

While organizational structure likely serves an important purpose, we believe that it need not necessarily be the created by the user, but could also be automatically generated by the system. Further, the organizational structure need not be static. Alex could issue a query for "football jersey", be reminded of any similar searches he ran earlier, and then use one of those similar searches to find the company, essentially using the search results like dynamic folders. Similarly, Yee, et al. [7] create dynamic organizational structure by allowing users to browse faceted meta-data.

Short term goal: Make advance work unnecessary for re-finding.

Note that while there is disagreement as to whether advance organization should be required of the user, none of the participants believe effectively being able to find information will make the process of organizing obsolete. Information organization furthers the user's understanding of the information space and helps the user remember the information being organized. In fact, people who file their information rather than pile it are more likely to use keyword search when looking for something [6], perhaps because of the role organization plays in helping them memorize and understand the information.

People should not have to do any work at all to re-find.

3 participants agree, 3 disagree

Just as catching a ball is only a part of a greater game such as football, so is re-finding, and, indeed, all of information management, just an activity that is part of a greater task. While in our original example, Alex was asked to re-find the name of the company he originally purchased team jerseys from, that name was necessary only because the team needed new football jerseys. While in this report we primarily discuss re-finding in isolation, it is important to consider the activity's greater context.

Ideally, a user would not have to do any work to re-find information at all. The previous section talked about doing away with the "tossing" in personal information management. The participants who believe that users should not have to do any work to re-find believe the "catching" of personal information

should also be done away with. Instead, relevant information should just appear when and where the user needs it.

Examples of this exist already in many small and task specific ways. For example, many email clients support filling in the recipient's email address as the user starts to type his name so that the user doesn't need to actually find the address. One could imagine even more clairvoyant systems that know, based on the user's context, the likely email recipient and automatically fill in the recipient field. Other examples of task-embedded re-finding include the Remembrance Agent [3], which re-finds documents relate to a document being currently composed, and Aria [5], which re-finds images related to an email being currently composed.

Another argument in support of having relevant information automatically appear in the context of the task is that we unanimously agree that information that the user doesn't remember having encountered can still have value. Such information is difficult to re-find, since the user does not even remember it exists. Pushing relevant information on the user could serve as an important reminding function.

Those who disagree with the statement, "People should not have to do any work at all to re-find," think that it is fine as an ideal, but entirely impractical as a solution. If a computer will never be able to perfectly guess the user's information needs, there will always be a need for information seeking tools. Thus, it is best put as a long term goal.

Long term goal: Make it so people don't have to re-find.

Re-finding is always part of another task. It's reuse that matters, not re-finding.

6 participants agree, 0 disagree

The participants were unanimous in their belief that the only reason to re-find information is to use the information target to accomplish some task of which re-finding is only a step. This assertion supports the argument above that users should not have to do any work in order to re-find appropriate information. That is, the ideal system is one that, in the process of accomplishing some task, is able to suggest information that the person has already encountered when it is needed, without forcing the person to leave the task of interest in order to engage in re-finding behavior. In order to achieve the ideal, it is important to understand the relationships between task types and potential information support for those tasks.

Medium term goal: Classification of tasks for which information support is important.

Pruning is good for re-finding. Support for losing is as important as support for re-finding.

3 participants agree, 3 disagree

There was considerable disagreement as to whether pruning information from an individual's personal information store would aid re-finding. The argument in favor of pruning is that people are currently subject to information overload, and do not want to have to interact with as much information as they do. By removing information from the user's information space, the user can feel more in control of that space and be better able to find important information nuggets.

Those that disagree with the statement, "pruning is good," believe that a good information management system can provide relief from information overload by *virtually* losing the information, while still retaining the data somewhere, should the user happen to need it in the future. Information can appear lost to the user without actually being removed from the computer.

A benefit to actually losing information that virtual loss would not provide is that it creates additional disk space. This is only a problem if disk space is constrained. It currently appears that this will not be a problem, but it could be an issue with the capture of large amounts of data about the user (e.g., continuous video feed of the user's life).

Conclusion

With respect to all of the assertions that we have posed, it is absolutely essential that there be a means by which they can be evaluated, and by which theories and techniques for understanding and addressing these issues can be tested. The most successful mode of evaluation to date in information retrieval research has been the use of test beds which allow many different investigations to be performed and compared (e.g. the TREC and MUC programs). Test beds for the evaluation of systems that support the re-finding of personal information will need to be substantially different from those that are currently available, since they will require knowledge of context, specification of task, and some record of interaction with information objects within task context. But such a resource would be enormously important for promoting real scientific research in this area, testing hypotheses, comparing approaches, and building on previous results.

Resources needed: An evaluation framework and methodology, and a testbed.

The validity of each of the above assertions is an open research question, and we throw this report into the future in hopes that researchers will catch it and be inspired to shed light on the statements.

References

1. Cool, C. and Belkin, N.J. (2002). A Classification of Interactions with Information. In *Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4)*, 1-15.
2. Jones, W. (2005). Introductory Remarks, *PIM Workshop*, Seattle, WA, January 2005.
3. Lieberman, H., Rosenzweig, E. and Singh, P. (2001). Aria: An Agent for Annotating and Retrieving Images. *IEEE Computer*, July 2001, 57-61.
4. Miller, M.J. (2005) Google, Yahoo!, and MSN: The Search Continues. *PC Magazine*, March 22, 2005, 5.
5. Rhodes, B. and Starner, T. (1996). The Remembrance Agent: A Continuously Running Automated Information Retrieval System. In *Proceedings of 1st International Conference on the Practical Application of Intelligent Agents and Multiagent Technology (PAAM '96)*, 487-495.
6. Teevan, J., Alvarado, C., Ackerman, M.S. and Karger, D.R. (2004). The Perfect Search Engine is Not Enough: A Study of Orienteering Behavior in Directed Search. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '04),* 415-422.
7. Yee, K-P., Swearingen, K., Li, K. and Hearst, M. (2003). Faceted Metadata for Image Search and Browsing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '03),* 401-408.

3.3 Encountering, keeping, organizing & maintaining information

Cathy Marshall, William Jones, Harry Bruce, Brian Ross, Tiziana Catarci, Doug Gage, David Maier

Although we recognize the tremendous potential of search, it is vital to remember that not all of the information that comes into our purview is actively sought to meet an established need. Information is encountered in the course of our everyday activities that may not be immediately useful. Rather, it may have anticipated value either as a reminder, for its evocative qualities, for its educational value, for the ideas it spurs, for its potential utility as a reference, or as something to share. Deciding what to do with encountered information, whether to keep and, if kept, how, form a key challenge of personal information management (PIM).

We encounter information in many different ways and forms. Sometimes we come across an interesting article when we're reading the news. A directed search may return an unexpected result, potentially useful in another context. A colleague may email us a URL or document. A forgotten photo appears when we explore an unlabeled disk. Even in an age of increased personalization, filtering, and ranking, we still have many serendipitous encounters and re-encounters with information in our everyday lives [0].

Fieldwork reveals that people keep information in many different ways and for many different reasons [1, 5]. Sometimes people will keep the same piece of information in two or three different ways "to be sure" of they can get back to the information again later and will remember to do so. People might, for example, bookmark a NY times article on the web, then also save this page to their hard drive (in case it disappears from web site), and finally email the reference to themselves so they will remember to look at it. The number of ways to keep and manage information has grown considerably in recent years, in step with the overall increase in the number of devices, technologies, and applications we rely on. The attendant fragmentation of our personal information increases the chances of keeping something in the wrong place or form.

Furthermore, encountered information may fall outside the usual assumptions that underlie PIM technologies. For example, we may find work-related information while we're at home, and vice-versa. We are often interrupting a task rather than performing one. And we may not yet have the appropriate filing structure to store the encountered information (other than the "misc" folder). . Encountered information may reflect *potential* interests – hobbies we haven't yet undertaken, projects we anticipate, trips we might take – and may not adhere to our current relatively well-conceived organizational habits, structures, and systems.

The capacities for digital storage continue to increase, making it possible for us to take a "keep everything" approach to the information we encounter (see the "Digital Memories" article in this issue). But our capacity to attend to information is not increasing in the same way [3]. Indeed, in field interviews directed at uncovering what people do (and hope to do) with the encountered material they keep, the term "pack rat" is often used to describe ineffective keeping strategies that cause valuable material to be buried by things indiscriminately kept [5]. Furthermore, both authors have observed that people often don't remember that they already have saved potentially useful or meaningful material when it might be brought to bear on the problem at hand. *You can't search for something if you don't remember that you have it in the first place*.

Moreover, we use the items we keep in ways that are not fully described by their searchable content [5]. What we keep may be emotionally evocative, reminding us of a place or event; we expect this material to stir memories through future re-encounters. Or, by contrast, what we keep may have a briefer lifespan as a visible reminder of what we plan to do: to remember to go to an art gallery opening or to try a new restaurant reviewed in the newspaper. But much of the material falls into a murkier middle ground of utility and permanence: we're not sure how long we need to keep the material and what exactly we'll use it for. The very act of keeping and organizing information appears to affect not only whether we remember the information but what about this information we notice and remember [4]. As such the acting of keeping may be a very useful step towards understanding the information better.

Thus much of what we keep represents a balancing act: the material must seem sufficiently useful, be sufficiently necessary as a reminder, be sufficiently compelling as a source of ideas, or be sufficiently evocative to merit the cognitive overhead of keeping it and the risk of mis-keeping it [3].

If information access and communication technologies have increased the amount of information we encounter and the fragmentation of what we keep, we may also look to tools and technologies for help. Good filters may already help by screening out junk email and deceptive web sites, for example. Categorizing tools may play a more positive role as well by helping us match encountered information to

areas of personal interest [6] We can also develop a more uniform infrastructure for facilities that help us highlight, annotate and set reminders to mark information for later use.

What broader implication does encountered information have for PIM tools? Our field experiences suggest that allowing material to accumulate and relying on search to reclaim it at the right time is insufficient; the sense-making activities that surround keeping are critical for later finding, whether they're associating material with a particular taxonomy or establishing a stable sense of place (e.g. Personal Unifying Taxonomies [0] or stable information geographies [0]). A good match between how something is kept and its envisioned role or function is essential. To develop effective PIM tools, it is important to remember that the utility, serendipity and pleasure of re-encountering what we have saved relies on more than search alone.

References

Bruce, H., Jones, W., Dumais, S. (2004). Information behaviour that keeps found things found. *Information Research*, Vol. 10 No. 1, October 2004. http://informationr.net/ir/10-1/paper207.html

Erdelez, S. Information Encountering: A conceptual framework for accidental information discovery. *Proc. International Conf. on Research in Information Needs, Seeking, and Use in Different Contexts.* Taylor Graham (1997), 412-421.

Jones, W. Finders, keepers? The present and future perfect in support of personal information management. *First Monday*, March 3, 2004. http://www.firstmonday.org/issues/issue9_3/jones/

Jones, W., Phuwanartnurak, A. J., Gill, R & Bruce, H. (2005). Don't Take My Folders Away! Organizing Personal Information to Get Things Done. Paper presented at the Conference on Human Factors in Computing Systems (CHI 2005), Portland, OR. (http://kftf.ischool.washington.edu/publications.asp)

Marshall, C.C. and Bly, S. Saving and Using Encountered Information: Implications for Electronic Periodicals. In *Proceedings of CHI'05*. New York: ACM Press (2005), 111-120.

Segal, R. and Kephart, J. MailCat: An Intelligent Assistant for Organizing E-Mail. In *Proceedings of the Third International Conference on Autonomous Agents*, May 1999.

3.4 From PIM to GIM

Facilitators: Thomas Erickson, Jonathan Grudin

Participants: Sue Dumais, Alon Halevy, Kent Unruh, Steve Whittaker

3.4.1 1. Preface

The charter of this group was to explore the question of how personal information management (PIM) relates to group information management (GIM). Motivating questions included how and where PIM fits in the business world, what happens to PIM when the information is not personally owned, and the implications of sharing personal information (intentionally or unintentionally).

As GIM does not exist as a recognized subfield or even phrase, the discussion began with an effort to agree upon a definition. The definitional effort quickly turned into an interplay between attempts to define a model of GIM, and the discussion of particular examples of group information management. As the group converged on a definition of PIM, attention shifted to a consideration of the problems and opportunities offered by GIM. Here, as well, there was an interplay between the discussion of examples and the articulation of problems and opportunities.

3.4.2 2. From PIM to GIM

Broadly put, personal information management (hereafter PIM) serves two ends, instrumental and symbolic. First, artifacts such as "to do" lists, calendars and rolodexes serve as external memories, and enable their users to efficiently conduct their daily tasks. Second, PIM can assist users in managing the impressions that others form of them. Thus, the use of a Day-Timer® or other personal organizer system —available in a wide array of materials ("an expression of your unique style," according to the Day-Timer web site)—can contribute to creating the impression of the user as an productive, well organized professional. And of course, if the PIM artifacts are adroitly deployed and succeed in achieving their instrumental ends, the impression will be augmented by efficiency of the user's performance.

PIM can also be seen as functioning in two spheres: private and more public. In the private sphere PIM simply supports one's personal tasks. Thus, while one's PIM activities may be glimpsed by others—as when we see someone checking an item off a list or looking up a number in a rolodex—the information is purely for the use of its owner. But PIM also functions in a more public sphere. That is, information is often created with some degree of sharing in mind. A student may take notes, writing a bit more carefully than usual, to share with an absent friend. Or members of a workgroup may develop a practice of sharing their calendars with one another to facilitate meeting scheduling. While this sharing serves the instrumental ends that motivated it, the information thus shared also becomes grist for possible inferences about the owner: the student's handwriting may be sloppy or her notes incomplete; a workgroup member's calendar may reveal private information such as medical appointments, or consistently long lunch dates. Thus, when personal information is shared, it introduces tensions between the instrumental ends for which it is shared and the not necessarily desirable inferences that it may support.

The tensions that occur as personal information is shared are complex and intertwined, and moreover have the potential to feed back and alter norms having to do with what is shared, and how it is shared. As a consequence, this area seems a valuable focus for research attention. We have adopted the phrase "Group Information Management" (hereafter GIM) to refer to PIM as it functions in more public spheres. More specifically, we define it as follows: *GIM has to do with how personal information is shared amongst a group, with an emphasis on the norms that underlie that sharing, and the ways in which participants negotiate those norms in response to a variety of tensions.*

3.4.3 3. Examples of GIM

Because GIM has to do with how information is shared amongst a group, it is not surprising that a wide array of applications can be used to support GIM, including email, web pages, WIKIs, and traditionally produced documents. However, while many applications can be turned to GIM ends, there are some that fall more squarely into the GIM arena.

3.4.3.1.1 Shared Calendars

An early example of GIM in the digital realm is the development of online calendaring systems. In the 1980's, various developers produced digital analogs of personal calendars that were designed to be shared by groups in an organization. The motivating idea was quite simple: making a person's calendar available to others could facilitate the sometimes onerous task of scheduling a meeting. However, these electronic calendaring systems encountered resistance for reasons ranging from the fact that personal calendars rarely contained a complete picture of their users' availability, to users' realization that calendar information could be used for other not necessarily desirable ends, such as making inferences about users' personal activities. As shared calendaring systems have been adopted and 'naturalized' within organizations, a variety of technical features and social practices have arisen in response to such tensions.

3.4.3.1.2 Blogs

A more recent application genre is the blog, a web-based, person-centric diary-like document consisting of relatively short entries displayed in reverse chronological order; these entries can be linked to, and commented upon, by readers. Most blogs are published by individuals for small audiences comprised of family and friends; however, some are published by groups for the explicit purpose of sharing information and generating commentary from a larger audience. Blogs raise interesting issues about audience: Who does the blog author imagine that he or she is writing for? What are the consequences when personal information published in a blog receives attention from a different audience than intended? What steps do blog authors take to avoid these consequences or recover from them?

3.4.3.1.3 Social Networking Services

Social networking services such as Orkut, LinkedIn® and Friendster® allow their users to post personal profiles, pictures, and create links to others signifying professional or social ties. The networks of links thus formed can then be viewed, traversed and used to distribute messages. Such systems serve a variety of purposes from supporting online professional networking to enabling singles to find prospective dates. Social networking services raise interesting issues about what users choose to reveal or conceal, how their disclosure of personal information is related to the ends that they hope to achieve, and the ethics of 'counterfeiting' links or conspiring to garner 'inauthentic' recommendations to increase their stature in the system.

3.4.3.1.4 Electronic Medical Records

As the information technology systems of the medical and insurance industries become increasingly interlinked, electronic medical records become an increasingly interesting example of GIM. Any particular patient's medical record is composed of information generated by multiple people (and devices); those who contribute to the record may come from different institutions, and enter information for a variety of different purposes. Access to records is by a similarly disparate audience for even more diverse purposes, and questions of ownership and access privileges are complex. This application of GIM raises complex questions of privacy and access and of ownership.

This list is not, of course, an exhaustive one. Other GIM-centered application areas include peer to peer file sharing, information sharing and tagging systems such as del.ici.ous and flickr, online reviewing and rating systems, and event organizing applications such as eVite® and MeetUp™.

3.4.4 4. Issues and Opportunities in GIM

After working towards a definition of GIM, and generating a list of core examples of GIM, the breakout group generated a list of research issues and opportunities that arise in GIM.

3.4.4.1 4.1. A Simple Model

To organize these issues, let's start with an overly simple model of GIM, and examine each part of the model:

A person generates information…

… that is shared with a group…

… in support of some task

Thus, an employee enters appointments in her calendar to share with her coworkers to facilitate the scheduling of meetings. Or a person creates a profile in an online social networking system to be shared with other members of the system for the purpose of getting dates.

A person generates information…

One set of GIM issues has to do with the creation of the to be shared information. What information do people choose to share, and why? (The implication of the model that they do so in support of a task, with a particular group in mind, is simply a conjecture, and, regardless, doesn't explain all cases.) What are the psychological issues that attend the decision to share information (for example, people have been observed to 'clean up' information before sharing it)? What are the various norms that attend sharing, and how do they vary according to form, content and domain of the information?

…that is shared with a group…

Another set of issues has to do with whom the information is shared. How is the audience for the information specified? How does the imagined audience interact with the nature of the information shared? What are the consequences of changes in the audience over time (for example, as an organizationally defined group changes composition)? What are the consequences of 'leakage' of the information beyond the intended audience? And, to the extent that GIM users are concerned with such questions, how might GIM systems support them in preventing or mitigating these issues?

…in support of some task

If we accept that in some cases people choose to share information in support of an envisioned task, what happens when the information turns out to be useful for other tasks that are not in the user's best interests? To what extent is it possible to give users control over uses of their personal information? To what extent is it possible to allow them to retract it after the intended task is completed? To what extent is it possible to simply allow users to be aware of when their information is actually used?

3.4.4.2 4.2. More Complex Variants of the Model

The above model is quite simple. Let's consider a few variations on it.

3.4.4.2.1 Ownership is complicated

The simple model assumes that personal information is owned by an individual. But in fact this assumption can fail in many ways. It may be that the individual is not voluntarily generating the information, as implied in the simple model, but is generating information as a side effect of his or her activities (e.g. credit information; medical records; calling records). In such cases individuals may not own their information, or if they own it may nevertheless lack complete control over its content, distribution or use. This raises a host of issues about who can see the information, how it can be used, whether it can be corrected if in error, or retracted if no longer needed, and how to deal with real or asserted errors in the information, or its distribution or use.

3.4.4.2.2 Information is generated collectively

In some cases groups may generate information collectively, creating either a single collective product (e.g. the contents of a Wiki), or a set of individual products that are shared with the other members of the group (e.g. the profiles in a social networking system). These cases are interesting because group norms and incentive systems come into play, and their establishment, support and evolution can play a critical role in shaping the character of the system. These effect the system at all levels, including the

nature of the information shared, and how collectively produced information is structured and the maintained over time.

GIM occurs in the context of an institution

GIM often occurs in an institutional context, and thus is shaped by institutional values, practices and mechanisms. Examining the practices of institutions that have developed expertise at GIM (e.g. the Mayo Clinic, with its century of experience in maintaining, sharing and glossing patient records) seems one fruitful avenue of exploration. Similarly, looking at the ways in which the needs of differing institutions (e.g. the medical and insurance industries) shape the nature and use of personal information also seems of interest.

3.4.5 5. Summary

GIM does not currently exist as a distinct field. Even the term, which occasionally occurs in the literature, mostly in the company of PIM, does not have an agreed upon meaning. As the discussion outlined in this document makes clear, GIM raises a number of interesting issues, and has considerable potential as a focus for research.

3.5 Measurement and evaluation

Diane Kelly, Wanda Pratt, Jim Gemmell, Mary Czerwinski

3.5.1 Creating Sharable Test Collections

We had a short discussion of the possibility of creating sharable test collections to study PIM. Our group was divided on the potential benefit and feasibility of such an effort, so we did not pursue this topic in-depth. We all agreed that creating a TREC style collection and using this collection to conduct interactive experiments where new subjects simulated tasks and personal information management activities was not a realistic or valid approach. However, we saw value in creating a PIM collection that other researchers could use to examine their own questions, techniques and applications. We acknowledged that the creation of sharable test collections can potentially facilitate discovery and allow for more rapid progress since building a good test collection is such a difficult, laborious, and time-consuming task. Standard test collections also allow for multiple modes of inquiry including those that involve the comparison of various techniques, examination of alternative hypotheses and replication of previous findings.

The design and construction of a test collection for PIMs would be an ambitious project. There are major issues related to privacy and generality. Clearly it would be necessary to identify what types of information should be included in such a collection. It would also be necessary to obtain subjects' permission ahead of time to make their data available to others and to clean the dataset to insure that sensitive information is deleted. Some kinds of data remain controversial – such as contact information of others or email received from others. Does the person whose information store contains this information about others really have the right to disclose it? Even if privacy concerns were addressed, the issue of whether a study of *personal* information can really be studied on someone else's information – which is essentially *not* personal.

3.5.2 Evaluation Design

One of the biggest challenges of studying PIM is that what we were interested in studying changes constantly. Furthermore, if PIMs is, in part, about "throwing information into the future," then what we want to study will happen at some unspecified, and usually unpredictable, time in the future. The nature of the information that we study poses further challenges. This information is personal and different for each user. Over time, users create their own idiosyncratic information collections and execute a wide variety of information management tasks and behaviors that are within the context of such collections. Finally, users' interactions with information objects are not discrete, and are very often dependent on their interactions with other objects. Given these challenges, we identified a number of

experimental designs that seemed most appropriate for the study of PIM. In general, we recommend mixed-method approaches, the use of both quantitative and qualitative methods, and triangulation.

Naturalistic, longitudinal approaches are very appropriate since these approaches allow one to capture data over on extended period of time and to take measurements at fixed points in time. These approaches also allow for users to conduct their natural information management activities and behaviors, in familiar environments, with familiar tools. One challenge of conducting a longitudinal study is the determination of an appropriate measurement interval. *When* you measure is just as important as *what* you measure, and this can vary based on what people are trying to accomplish at any given moment in time. Further, a person's activities and behaviors are often governed by external events, which can impact what kinds of PIM you are likely to observe. Case studies, which focus attention on one or a few users, are also valuable approaches to studying PIM. Case studies most often produce rich, descriptive results, which in addition to being important in their own right, can also lead to explanatory studies. Although intensive approaches to data collection do not usually allow one to study large samples, the quality and quantity of data that one gathers about a small number of users can be quite extraordinary. While this data may not be representative of the behavior of a larger population, this data is much more representative of those users' behaviors. These types of approaches further optimize the ecological validity. However, caution regarding overgeneralization from too few cases is something of which to be mindful. A very practical concern that we have is publishing research of this kind; many publishing venues look more favorably toward research with large numbers of subjects.

We also identified value in using laboratory studies to investigate PIM. Given that laboratory studies involve a great deal of reduction, it is important that what is being studied is a bit more defined and narrow in scope. For instance, it does not make much sense to study general PIM in a laboratory setting. However, leveraging the power that laboratory studies offer is definitely something that needs to be included in any evaluation framework for PIM. One prevalent challenge to conducting laboratory studies is simulating users' real-world use environments. In particular, the collection is an important consideration, as are tasks which users are asked to conduct. Most laboratory studies involve a known, general collection of information; asking users to conduct PIM tasks with collections about which they know little (or nothing) raises some validity issues. One compelling suggestion is to ask users to provide their own information collections. However, this requires users to do more preparation work, and further, users will likely be very selective about what they include in their experimental collection. Control is also issue; some users may prepare a collection of 5,000 photos, while another may prepare a collection of 50 photos. Task is a bigger issue, which we address below.

We agreed that traditional experimental designs might offer our community a basic framework for conducing evaluations. In particular, designs that include a [pre-test | treatment | post-test] might offer a promising approach. The Solomon Four Group design, which is rarely used in any of our fields of inquiry, might also provide an interesting perspective on PIM evaluation. Examining the methods used in fields that employ this type of design, such as Education, might assist us with identifying potentially useful approaches to studying PIM.

Establishing a baseline to measure changes in behaviors poses a significant challenge to PIM research, since everyone's baseline is different. In many ways, this implies that we will need to assess individual baselines before our study commences and measure changes with respect to each individual. Ethnography and observation might be ways to assess these individual baselines in a naturalistic setting.

A typical approach to collecting data about users is to collect log data, and we feel that this approach is certainly relevant to PIM. Studies based exclusively on log data are attractive since a great deal of data can be collected in a relatively short period of time. However, caution must be taken when relying too much on log data, since log data necessarily represents an incomplete picture of a user's activities. For instance, log data does not tell us about a user's goals and intentions. Further, it is of utmost importance to make sure that one's log data is valid and reliable. If it does not meet these basic criteria, then it is worthless.

The ability to do rapid prototyping and deployment is also an issue that we discussed. A PIM tool could take several years to develop. How can we use rapid prototyping to quickly get a tool to

users? The development of stable, usable PIM tools presents a challenge for us all and we are in much need of a framework for rapid prototyping and deployment.

3.5.2.1 Tasks

One evaluation issue that we spent some time discussing was the issue of tasks. The types of tasks that are relevant to PIM are very broad, user-centric and situation-specific. Further, tasks are often identified at varying levels of specificity. For instance, "doing email" is a task, but one might subdivide this task into searching for a specific piece of email, managing and filing emails, setting up an address book, etc. We feel that there are many generic classes of tasks that users do, such as "finding information about X," "reading the news," and "planning travel." However, in a real environment there is no way to anticipate the number and kinds of tasks that users are doing. Tasks also differ according to the length of time they take to accomplish and the frequency with which users work on them. We discussed the idea of multitasking as a way of life and that good PIM should support seamless task switching and integration of activities. In many cases, users abandon tools because the tasks that they are meant to support are so short and occur with such frequency that opening a new application is too much work. Instead of thinking of singular tasks, perhaps we should develop sets of tasks for laboratory studies to simulate multitasking behavior.

Finally, re-finding tasks present a unique challenge because it is *use* not *search* that is the goal of re-finding tasks. Sometimes re-finding a piece of information is not good enough because the information lacks clues about the original context of access/use. Without this type of information, it is often difficult for users to understand their original interpretations and intentions behind viewing the information in the first place.

3.5.3 Users

We all agreed that many of the evaluation designs that we identified dictate the use of a small number of users. In studies where there are a small number of users, we recommend that much effort be spent profiling users. We identified several characteristics that are important to gather about the user: age, sex, ethnicity, experience (search and otherwise), education, and various cognitive abilities (e.g. spatial/intellectual/motor abilities). Continuing to identify best practices for profiling users is an important topic for future study.

3.5.4 Measurement

Our group spent quite a bit of time discussing measurement. Measurement has to be understood within the context of some particular PIM goal. What are the goals of PIM? What are the effects of PIM? What is PIM suppose to help us accomplish? How do we know good PIM when we see it/experience it? How do evaluate whether [more | less] of something is [good | bad]? Several measures were identified during our session. In general, we agreed that subjective and affective measures are important and critical. We also discussed the use of indirect or implicit measures such as quality of life assessment and improved decision-making as indicators of how PIM impacts people's lives or changes their behaviors.

3.5.4.1 Efficiency & Time

Efficiency is an interesting measure because 'good' PIM might actually allow a person to spend more time on certain types of tasks, rather than allowing for the completion of more tasks in the same amount of time. Not only can PIM allow potentially for more tasks to be done in the same amount of time, PIM can also allow for fewer tasks (or the same amount of tasks) to be done better. At a minimum, efficiency measures need to consider time, and quantity and quality of output. Some other questions that we asked with regard to efficiency: Are you checking more things off your 'to-do' list? Do you spend more time on your high priority tasks than you used to? Do you spend less time on your low priority tasks than you used to? We do not recommend centering one's evaluation on whether or not a novice user can learn to use a tool in five minutes.

Re-finding tasks present a special case of using time as a measure of success, since use is the ultimate goal of most re-finding tasks. As described earlier, one has to re-find information before one can use it; thus it seems more appropriate to consider the time it takes someone to formally use the

information (e.g. including it in a report) rather than the time it takes someone to locate or find the information, as a measure.

3.5.4.2 Flow

We agreed that good PIM ought to allow people to be 'in the flow' when they work and to concentrate on more important tasks. In particular, PIM might decrease flow if people have to waste time filing for future activity instead of focusing on the task at hand. It is critical that our PIM tools are integrated seamlessly into our day-to-day activities and are not just another distraction. Currently, there are three proxies for flow: relative duration, general satisfaction and happiness, and cognitive function (ability to do the work even when there are external cognitive distractions). A person's ability to perceive these external distractions and interruptions might also indicate flow. Presumably, if one is in 'the flow,' then one should be able to ignore distractions and be able to accomplish complex tasks and activities.

3.5.4.3 Use

Use is a measure that can indicate a great deal about the value of PIM. The behavior of adopting a tool and incorporating it one's life can be considered as one indicator of value or success. Repeated use is a good indicator of success. Understanding how many people do not use your tool (or abandoned your tool) can also be a good metric. However, taking a simple measure of use/disuse/abuse(?) is limited since we are unable to understand what does and does not work, and why a person has adopted (or not adopted) a tool.

3.5.4.4 Quality of Life

An interesting set of measures that has received little attention in most of our disciplines is quality of life measures. Quality of life measures could act potentially as indirect measures of the success of PIM. One generally agreed upon goal of PIM is to make our lives easier and to perhaps free up some of our time so that we can enjoy a variety of life experiences (not just work!). Quality of life measures can allow us to potentially understand the broader impact that PIM is having in our lives. Wanda Pratt called our attention to the following quality of life questionnaire, which might be used as a starting point to the integration of these types of measures into our research:

Endicott, J., Nee, J., Harrison, W., & Blumenthal, R. (1993). Quality of life enjoyment and satisfaction questionnaire: A new measure. *Psychopharmacol Bulletin 29*(2), 321-326.

3.5.4.5 Process/Behavioral Changes

Success, in part, can be viewed as making a positive change in process or making a positive change in a person's behaviors. For instance, a positive change in a person's decision-making ability as a result of PIM, is a good indication of value. The difficult part is isolating variables in order to demonstrate cause and effect. Given the complexities of our work environments and idiosyncrasies in our behaviors, this is a serious evaluation challenge. Another potential measure of process change with respect to a group work setting is worker productivity improvement. In these types of situations, objective raters might be used to evaluate the quality of the group work and each group member can be assessed individually, by fellow group members.

3.5.4.6 Subjective Duration Assessment

Subjective duration assessment asks people to estimate the length of time it took them to complete a task and then compares this estimate to the actual length of time it took them to complete the task. The theory is that if a person underestimates the time, then the task was easy (and perhaps enjoyable) to accomplish. If a person overestimates the time, then the task was difficult (or the person did not finish). Accurately estimating the time indicates that the task was neither easy nor difficult. The value of subjective duration assessment is asking people to make estimates or predictions about their own behaviors in situations where you have the actual, objective measure with which to compare these estimates. These types of measures seem particularly applicable to web tasks and re-finding tasks.

3.5.5 Privacy

A final theme that was prevalent throughout our discussion was privacy and, more generally, ethics. Given that we are studying personal information it is worth reexamining our ethical obligations to subjects. It is also worth examining privacy issues which emerge as a result of the kind of information that we study. For instance, it is common to obtain permission from a subject to examine his/her email, but it is uncommon to obtain permission from each person who has sent that person email. In another example, consider the situation where one is investigating PIM in organizational settings. A hierarchy of privacy might dictate that a manager gives a researcher 'permission' to study his/her subordinates' personal information even though the subordinate is comfortable with sharing this information.

3.6 Towards a unification & integration of PIM support

David karger, William Jones, Ofer Bergman, Wanda Pratt, Mike Franklin

Information fragmentation is a pervasive problem which is felt in several stages of personal information management (PIM). As the example in the introduction to this special issue on PIM illustrates, even a seemingly simple decision, such as whether to say "yes" to an invitation, often depends upon a number of different kinds of information – information from a calendar, from a paper flyer, from web sites, from a previous email conversation, etc. Information can be fragmented by physical location. This is nothing new. Now information is often fragmented by the very tools that have been designed to help us manage our information. Our information may be scattered across various computers and gadgets. Some information, for example, may be on a laptop computer we use at home, other information may be on a desktop computer we use at work and one or more PDA or smart phones. Even on a single computer, our information is scattered across the computer desktop, "My Document", file folders, email folders, collections of bookmarks, etc. New applications such as Microsoft OneNote introduce still more forms of organization with little or no integration to previous forms. People can rightly complain that they have "too many hierarchies" and people sometimes go to great lengths to bring their information together into a single organization whether based in files, paper or email messages.

Information fragmentation creates problems not only in the maintenance of several organizations but also in everyday PIM actions such as keeping and finding. We may sometimes need to look in several places, physical and virtual, in order to gather together the information we need for a particular task. We may also be less certain where and how to keep newly encountered information. Or do we "have it" already? If we keep the information again anyway ("just in case") we may then face some serious problems with consistency and updating later on.

Even when information is not directly copied, today's applications often force us to repeat the same data in several places. For example, a name such as "Jill Johnson" might appear in an address book, and also as the creator of a photograph in a digital photo album. Changes to one version of the data (a new married name, for example) often do not propagate to other versions of the data. Also, we may experience the frustration of having some operations – name resolution, for example – available in one place (when sending email) but not in another (when working with photographs). Finally, there is no easy way to "link" together the various bits and pieces of data relating to "Jill Johnson". In some cases, we may need to perform a difficult search in order to access another representation of information we are already looking at!

If the computer has been an unintended agent of information fragmentation, it can also be used to help us "put the pieces together" again. This article provides a sampling of some of the ways in which our personal information might be better integrated. The article concludes with a look at three research prototypes that illustrate varieties of approach to the fundamental challenge of personal information integration.

3.6.1 Motivation: Variations in Unification

There are different kinds of unification, each with associated benefits:

Unification across physical location. Perhaps the most basic kind of unification is the unification of information from many physically distinct sources. It is a significant burden to move physically from location to location to get the information we need especially when these are separated by some distance. Computing technology, in several ways, has done a great deal to integrate information across physical location. Data transfer protocols such as FTP (for data transfer) and X windows (for display transfer) have long existed to bring information from where it is stored to where a we, as users, need it.
More recently, tools such as network file systems (NFS) and the Web free us from even having to think about the physical location of the desired information. As the capacity and portability of storage devices continues to increase we can now bring with us a substantial proportion of the information we use regularly – on a laptop, for example, or even on much smaller device such as a PDA or an Ipod. We can access still more information via a wireless connection. Moreover, we now have access to many kinds of information in digital form – text, of course, but also pictures, music and even full-motion video. Computing technologies already combine to enable a high degree of unification with respect to physical location.

Unification across forms of digital information. Several studies suggest that people would like a greater unification across forms of digital information as well. In particular, people express, in various ways, a desire for a more uniform treatment of digital documents, email messages and web pages – especially when these all relate to a single activity. A unification across forms of information has several different aspects: Information can be unified with respect to access routes and with respect to means of grouping, viewing and manipulation. Each of these is separately explored below.

Unification in access. Even as we use computing technology to cross large gulfs of physical location, we continue to struggle with a fragmentation created by the many digital organizations of information that often co-exist even on the same computer. When different applications -- such as our file manager, email client and web browser – manage separate organizations of information, we may need to perform essentially the same basic retrieval actions repeatedly in order to find all the information that relates to a given activity.

Some important steps have been taken to a more integrative access to digital information. For example, many desktop search utilities now support integrative searches that cross organizational boundaries in order to return, in a single listing, email messages, files and web pages, that match a user's query. Support for integrative searching is now finding its way into the operating system in new releases of both the Macintosh OS and Microsoft Windows.

However, studies continue to show that people have a strong preference for browsing or "orienteering" styles of access to their information. People use search only after these preferred methods of access fail. User-created folder structures provide one means to browse to information. Folder structures may provide other valuable functions as well. In one study, for example, user-created folder structures sometimes appeared to serve as a problem decomposition or project plan. People also reported that their folder structures gave them an important sense of control over their information and helped them to "see" their information better. But, folder structures also continue to separate information by form – files go into file folders, email messages into email folders, web references into folders accessed through a web browser. To be sure, people can decide to force all their information into a single organization – for example, by encapsulating documents as attachments in email messages which are then stored in an email folder structure – but this is extra work and many useful file system features are left behind in the process.

Even within the same folder organization competing organizational schemes may suffer an uneasy co-existence with each other. People may apply one scheme on one day and another scheme the day after. There may, for example, be a tension between organizing files (images, articles, etc.) by project

for current use and organizing these same files by content for repeated re-use . Although research indicates that, given the right support, people are able to assign multiple categories to an information item, the support available to the average computer user (e.g., for creation of shortcuts in Microsoft Windows or for aliases on the Macintosh) remains quite primitive.

Unification by grouping and association. A second kind of unification across digital forms of information is accomplished by better support for the grouping and interrelating of items – both to each other and to tasks for which they are needed. We might, for example, want to interrelate all the information for a particular person in our lives. Frequently, we group together information relating to a particular task we wish to complete. For example, we might group together information concerning hotels in a city in order to select a hotel for our stay in that city. Traditional folders provide one means for grouping information together. Research has explored the more general and flexible notion of a collection. Items can be manually assigned to a collection (e.g. files placed into a folder) but items (or at least suggested items) for a collection can also be generated based upon a match between items and a "definition" (e.g. a query) for the collection. Limited support for the automated creation of collections is available now via features such as Microsoft Outlook's "Search Folders". Variations on this are now expected in new releases of both the Macintosh and Microsoft Windows operation systems.

It is often useful to assign properties to a collection as a whole. For example, if a collection of information relates to a task ("Find a hotel"), then it may be useful to assign task-like properties like "remind by" and "due by" which might then appear as appointments in an electronic calendar or trigger a reminder (via pop-up or email message) later on.

Associations to various aspects of the current context are also a potential basis of unification. The time of our last interaction with a document (email message, web page) is recorded currently. But many other aspects of the interactive context are not. For example, as we create a new document, send an email message or navigate to a web page, we may have a particular task in mind, but there is very little support communicating this task to the computer. Newly created documents, for example, are often placed, by default, in a place like "My Documents". In general, the context we "share" with the computer in our interactions with information items is very limited.

Unification by view. A third kind of unification of digital information takes us inside a collection of information (however defined or created). We seek to "view" the items within a collection. We look for recurring patterns among and important connections between information the items in view. For paper documents, the desktop and other flat surfaces of an office traditionally serve as a view space. We may move paper documents from filing cabinets to the desktop in order to "see" the information better. Computers provide several alternatives for comparable viewings of digital information including the computer desktop, a folder listing of files (or email messages or web references) and the window displays of opened documents, email messages, web sites. As cognitive psychologists know, our view of items can act as a powerful extension to our limited internal working memory for information.

Unfortunately, as we attempt to arrange information on a computer display, we experience problems. For example, applications involved in rendering the items of a collection may each consume a large part of the display. Documents, email messages, web pages, etc. may each "live" in a large window with attendant menus, toolbars, jumping-off points and default presentations. Because the window manager treats the
application opaquely as a rectangle full of pixels, it cannot select
the one piece of the display that the user actually cares about.
A common consequence is *window clutter* as evidenced by a display filled with windows, often obscuring each other and each competing for our attention. We can experience similar problems with the computer desktop and, of course, with top of a physical desk. The information we lay out in order to see and understand can turn into a jumble that actually impedes our ability to work effectively.

Current computing support for the creation of more workable, integrative views of information is quite limited. There has been little progress in file managers, for example, beyond the standard icon, list

(possibly with properties) and thumbnail views. One problem is that very little data is readily available concerning a "file" for use by the file manager in making decisions concerning display. More generally, better support for the creation of integrative views of information depends, in part, upon having a richer foundation computer-usable data for files and other information items.

Unification in the facilities of data manipulation. In a fourth kind of unification, we move from "read" to "write" access. For example, we may want to give explicit, external representation to the patterns we notice, the connections we make and main points we note for information that we are viewing. Or we may want to transfer information from one application to another.
A basic facility of data manipulation that we use repeatedly in a typical day is the copy/cut & paste facility (and the drag & drop facility). The c & paste facility provides an intuitive way of moving data from one application to another – although in some cases, the transfer is still text-only.

Other facilities for manipulating data are still provided in a very fragmented, piecemeal fashion. For example, in ways that are analogous to those we use when marking up a paper document, it is possible to highlight and annotate selected text for a document in Microsoft Word. Similar, but not identical, operations can be made on "PDF" documents displayed in Adobe Acrobat. However, it is not possible to perform comparable operations on the selected text of an email message displayed in Microsoft Outlook nor is it possible to highlight the selected text of a Web page presented in the Microsoft Internet Explorer. Even the basic ability to impose an ordering on information items is unsupported (e.g., for email messages) or accomplished only by a clever use of leading characters.

This discussion is intended neither to be a definitive nor exhaustive treatment of the ways in which we might like to see a greater unification or integration of our information. However, the discussion should provide a sense for the many facets of a general term like "unification" as applied to PIM.. These facets (and others that may occur to the reader) can be used as a basis for comparing approaches to the unification of personal information. Three such approaches described in the next section.

3.7 Enhancements of personal information

Daivd Maier, Alon Halevy, Marcia Bates, Ben Bederson, Harry Bruce

Personal information as initially encountered can often be very raw, fragmentary, or partially relevant; it may come from disparate sources with differing format and structures. Hence there have been many proposals to enhance it in various ways to make it more useful for the task at hand, to improve later findability, or to record and reuse human analysis and judgment connected with it. Enhancements typically involve adding more data to personal information or adding links between previously unconnected pieces of personal information, but can involve deletions or removal of extraneous relationships. This breakout group discussed the kinds of enhancements that have been considered (or should be considered), the variety of reasons for enhancing personal information, and the issues that arise in devising enhancement methods. We recount our discussions on each in turn.

3.7.1 Examples of Enhancement

Enhancements to personal information can be done on individual pieces of information, or on collections of data. They can be performed (semi-) automatically by some system or manually by a user.

We first consider examples of enhancements to individual documents. The most common example of an enhancement to a single document is to annotate it. For example, we may add some annotations on the content of a particular photo, or comments on the context of a particular file in our directory. Judgments and rating of documents can also be useful for later recall and ranking in searches. Annotations can be attached to a document as a whole, or to some passage or other information element within a

document. An annotation may be as simple as highlighting a phrase, or as complex as interlinear markup of a foreign text with pronunciation, literal translation and idiomatic translation. Summarization is also a form of enhancement. Summaries of personal information items can later provide us quicker ways of recalling their content. Another important example of enhancement is providing the lineage or provenance of a particular document (or internal element of a document). The lineage may point to the document from which it was derived, the method by which it was derived, when it was created or modified, the instrument it was captured with (e.g., camera), or other activities that were being performed in parallel with the one relevant to the document at hand. Of course, cleaning data items (e.g., fixing name spellings) can enhance their quality later on, as can stripping headers or formatting that are no longer relevant.

A second class of enhancements involves adding links between related items of personal information. One example is linking the entry in one's contacts file to emails from the particular person, or even to the papers co-authored with that person. Such annotations enable use to easily cross application boundaries when we browse our personal information (in the spirit of the Memex vision). The group speculated on linkages between digital items and physical artifacts, so that, for example, rearranging sticky notes on a wall reorganized corresponding elements in an outline, or corrections written on a paper printout are reflected in the electronic version of a document.

Enhancements that involve collections of information can come in at least two varieties. The first is clustering data items in a semantically meaningful way. For example, grouping one's email into different *activities*, or clustering all the photos that involve a particular person, place or event. A second class of enhancements has the goal of improving the efficiency of locating particular data items. These will include novel indexes on the data or access methods for browsing the data (e.g., a new "virtual" directory structure that brings together information from multiple other directories). Adding a glossary of definitions used in a particular data set can also significantly enhance its readability in the future.

Finally, a more exotic kind of enhancement endows data items with dynamic capabilities, giving additional functionality beyond the provided by the applications for creating or viewing them. As one example, you may consider using data items as input an engine that selects advertisements that are likely to be relevant to you based on your personal information. A different kind of example is to add computational power to email. For example, an email message request votes or selections from the receivers, and provide capabilities for the receivers to vote and to automatically tabulate their replies. One can also imagine a module that hides or selected portions of a document depending on who is viewing it.

3.7.2 Reasons for Enhancing Personal Information

There are multiple reasons for enhancing personal information.

1. **Improve the quality of the information:** Information as received is often raw or unvetted and needs to be enhanced to make it more easily accessible and understandable by its users. Examples include cleaning data, excerpting from a web page, summarizing a document and trimming headers or inclusions from an email message.
2. **Reminding qualities:** Enhancements to personal information can help the user remember what they were doing when they first inspected or created the data, thereby providing context for (re)locating other relevant data.
3. **Efficiency:** Adding index structures enables more efficient search into our personal information. A different form of efficiency is to provide additional access paths to data by creating virtual directories (or other forms of super-imposed structure) that can support browsing or navigation.
4. **Add missing information:** It often happens that certain information is simply missing or would simply not be part of the application originating the data, such as annotations on photos or

lineage information. Similarly, adding links between disparate data items fills in gaps in our personal information, such as connecting two email addresses used by the same person.

5. **Repurpose the information:** Our personal data can often be leveraged for different tasks than were originally intended, and it is often necessary to enhance the data so it is relevant to other tasks (adding keywords or ratings, making document structure explicit, selecting subsets).
6. **Record results of human analysis:** an obvious reason to enhance personal information is to record the results of machine or human analysis and then carry them forward with the data (e.g., cleaning references to people or articles in one's personal information, marking questionable sections of a web page, flagging useful functions in a user guide).

3.7.3 Issues for the Mechanics of Enhancements

We discussed several issues regarding the implementation of enhancement modules and architectures for incorporating them. The following are some of the issues that came up and suggested principles for building enhancement tools. Note that not all of these are relevant in every enhancement context.

1. The enhancements should be available with the data later on – i.e., they "move with the data."
2. We should be able to access or reconstruct the original (un-enhanced) data.
3. Addition of enhancements should link data items across different media, and should not end at the edge of the screen.
4. Enhancements should be optional: Their creation should not necessarily be completely automated, and a user should always be able to reject a proposed enhancement.
5. Like the data itself, the enhancements should be searchable.
6. Enhancements should consider the internal structure of documents (e.g., spreadsheets, email messages), in order to associate with the appropriate granularity (column, cell; header field, body paragraph).
7. There is a need for a global scope for information beyond the actual data elements themselves. For example, we need to be able to store a link between a paragraph in a PDF file and a row in a database, even though the link is not stored with either of these data items.

3.8 Search, finding, filtering and auto-classification

"*Searching, Finding, Filtering and Auto-Classification*"

Breakout Group Report

Participants:

Facilitators: Nick Belkin, Susan Dumais (report author), Diane Kelly

Scribe: Luna Dong

Participants:

Nick Belkin (Rutgers University), http://www.scils.rutgers.edu/~belkin/belkin.html

Rick Boardman (Google), http://www.iis.ee.ic.ac.uk/~rick/

Susan Dumais (Microsoft Research), http://research.microsoft.com/~sdumais

Luna Dong (University of Washington), http://www.cs.washington.edu/homes/lunadong/

Jaime Teevan (MIT), http://people.csail.mit.edu/teevan/

Diane Kelly (University of North Carolina), http://www.ils.unc.edu/~dianek/

Brian Ross (University of Illinois), http://www.psych.uiuc.edu/people/showprofile.php?id=7

Overview:

We spent most of our time discussing how to help users re-find information in rich personal stores. Personal information comes from many different sources (email, files, web pages, calendar appointments, instant messages, rss feeds, newspapers, notes, music, images, videos, etc.), in many different formats, and in the context of many different primary activities (writing a paper, creating a presentation, organizing a meeting, reviewing a technical paper, planning a trip, catching up on email, reading the latest news headlines, etc.). People may organize the information into directory or folder structures, they may add annotations, or they may do nothing and rely on full-text search to find it again. The ability to handle the diversity of information types and metadata quality is critical in accessing personal information.

Looking for information in a personal store is different in many ways from a search in an unknown collection like the Web. Perhaps the most important difference is that people are familiar with many different characteristics of information as well as the contexts in which they have previously encountered it. Because people remember different characteristics of the information they are looking for at different times (e.g., who sent an email, when you created a document, the topic of a memo), it is important to support a wide variety of access routes. The idea of fast and flexible access to personal digital memories was popularized by Vannevar Bush in his seminal paper in 1945. Although the technologies are quite different than those envisioned by Bush, the latest operating systems (e.g., Apple's Tiger OSX and Microsoft's Vista OS) and new desktop search tools (e.g., Copernic, Google, HotBot, Lookout, MSN, X1, Yahoo) provide the infrastructure to support some key of his vision. Key challenges remain in combining automatic and human organization, and in providing interfaces to help users specify their information needs and understand the results returned.

Description:

We first considered characteristics of human memory, including what cues people remember about information and what kinds of initial processing will be most useful at retrieval time. We then turned to the topic of how to design systems that can better support people in organizing and harvesting personal memories.

Cues for search.

What do people remember about their personal information?

- Content. What is it about? This is a primary search cue on the Web and will be important in personal information as well. Because people have previously interacted with (created, read, modified) information in their personal stores, they will also remember a wide range of other cues about these items. Techniques for supporting access along many different dimensions, as well as tools for capturing metadata about objects (either automatically by recording attributes like time, or manually by allowing people to file or annotate items) are important to develop for PIM applications.

- Context. What was I doing when I encountered the item? What happened just before or after? How similar is the retrieval context to the context at time of previous encounters?

- Time. When did I initially encounter it? When did I subsequently use it?

- People. Who was involved?

- Storage location. Where did I file the item?

- Physical characteristics. What does the item 'look' like? This includes physical characteristics of the item (e.g. size, position, type, font) as well as the context in which I encountered it (e.g., other items that were around, ambient music, time of day, location).

- Distinctiveness. How distinct is this item? Slameka and other psychologist have shown that items which are distinct are easier to retrieve. The distinctiveness of duration, location and attendees are important in predicting which electronic calendar appointments people will find memorable (Horvitz, Dumais and Koch).

- Encoding effects. What did I do with the item? How items are processed when they are initially encountered has a large effect on how easily they can later be retrieved. For example, items that are processed more deeply are retrieved more easily (Craik and Lockhart). This might have interesting implications for automatic vs. manual filing of email. Manual filing (and other types of explicit organization) should improve the memorability of the item, but at the cost of additional processing time.

- Recency and frequency. Two important factors in retrieval from human memory are how often an item has previously been encountered (frequency and the spacing of practice) and when it was encountered (recency). Anderson and Schooler have argued that these characteristics of human memory can be considered as a rational adaptation to statistics of the item encountered in the world. From an information systems perspective, this suggests that temporal and usage factors should be incorporated into access schemes.

- Recall vs. recognition. It is much easier for people to recognize items from among a set of alternatives than to recall or generate the items. Retrieval of personal information is an interesting case, which may lie between these extremes. When people search for information they have seen before they remember some attributes, with varying degrees of accuracy. The variety of attributes that are stored when an item is encountered and how the information is presented will both improve this kind of cued recall.

There was some discussion of the extent to which searching and browsing are qualitatively different activities or are extremes of a continuum, but we did not resolve the issue. There was agreement that people do both and that they need to be able to go back and forth between them easily. We also discussed whether queries could be thought of as data, again an issue which we did not resolve. A common theme of our discussion was the need to support access using a wide range of cues that people might remember about items of interest.

Harvesting personal memories.

What kinds of aids can we provide to people to make it easier for them to access their personal memories? We broke this problem down into two main areas: communicating information needs and understanding the results that are presented.

C*ommunicating information needs.* Today the most common way for users to specify their information need is to type keywords into a small "search box", or browse a hierarchy of folders organized along a single attribute (typically folder name). We discussed a number of alternative techniques. Relevance feedback, in which people mark some items as relevant, is a well-know technique for improving the relevance of items in batch mode evaluations (Salton and Buckley). Interestingly there are few examples of its use in operational systems. Encouraging people to say more about their information needs has also been shown to improve retrieval accuracy in laboratory studies (Kelly). Spelling correction is a simple technique that works well when people misspell what they are looking for, although even here the details of how and when alternatives are presented has a large influence on the success of the approach (Mayer). Tabbed completion is another alternative that has been successful in some settings, although it is difficult for novices to discover. Recommendations are another technique for suggesting related items or query terms that might also be of interest. Capabilities that support richer interaction are also possible, including specification by selecting regions of the current

document, or richer representation and use of facets. Implicit queries can also be generated using current document contexts.

Understanding the results presentation. Today most retrieval systems return a long list of results which typically includes a title, url and short contextual description. Several experimental systems have explored alternatives, but few are widely used. Some systems have presented richer summaries such as thumbnails, query-relevant thumbnails, or additional details on demand. Others have grouped the results in some way, e.g., by site or by content using text clustering or text classification techniques. The use of richer faceted metadata has been explored by several groups, and appears to be especially important for personal information retrieval since people remember many attributes of items that they have previously encountered (e.g., Dumais et al.). Some metadata can be automatically captured (e.g., the time an item is receive, author, recipient, subject, interaction history, etc.), but we need to support users in specifying additional metadata as well. Popular folder structures are one type of metadata, but others could include things like a "keeper button" that allows users to mark the current item as important or to save the current context for subsequent presentation. There is a tight coupling between storage and retrieval and we need to consider both in designing systems. There was an interesting discussion of the extent to which the same cues are useful for finding and re-finding.

Key research challenges:

There are tremendous opportunities to go beyond the popular search box and a long list of results to help people to specify their information needs and to understand the relations among results that are returned.

There are large individual differences (both across individuals and within an individual for different task) in the strategies for organizing and retrieving information. Understanding the costs and benefits of investing in saving and organizing information is an important first step – i.e., to what extent are the costs invested in organizing information worth it in terms of retrieval accuracy or speed; can we develop tools to mitigate costs and improve benefits? The evolution of content, strategies and access patterns over time is an important dimension that is just beginning to be explored.

The ability to handle diverse types of information and metadata is critical for accessing personal information. People create and encounter many different kinds of information and it all needs to be accessible. People remember many cues about information they have previously encountered and systems need to provide multiple access routes and allow users to switch easily between them. Iteration and interaction, rather than one-shot querying, need to be supported. Richer visualizations showing the relationships among retrieved items might be appropriate for some information analysis applications.

It is also important to develop new techniques for specifying information needs that go beyond a simple search box. People bring much more to retrieval situations than 2.5 words, including rich search histories. Researchers should work to develop techniques to elicit and capture this information and incorporate it into retrieval. Understanding what and when to elicit (either explicitly or automatically) are important topics for future research.

Information retrieval is driven by information needs, so a richer understanding of users' tasks and contexts is central to developing systems that support the management and retrieval of personal information.

Final thoughts and a parting image:

The image below is a screen shot of the final poster our group used to summarize our discussion at the workshop. It might not be memorable to readers, but to those of us who were involved in its creation, it will serve as an important memory cue for years to come!

3.9 Digital memories, ubiquitous computing

Mary Czerwinski, Jim Gemmell, Doug Gage, Cathy Marshall, Tiziana Catarci, Manuel Perez

3.9.1 **Scenario**

Gordon returns from his business trip and the photos taken automatically from his hat-mounted camera (Figure 2 - left) begin to appear on the screen saver of his fridge (Figure 2 - right). One picture of a lunch with colleagues reminds him of an email message he wants to look at again. He knows he read the email during a meeting after the lunch. On his tablet PC, he opens the list of photos played in his screen saver, and looks up the appointment associated with photo. He sees the subsequent meeting, and requests all emails accessed during the meeting.

After finding a reviewing the email he wanted, he decides to share some photos of the trip with friends. He wants to find a particular photo, and the first thing he remembers is that it was a very hot afternoon. So, he searches for all photos taken when his personal sensor read more than 80 degrees. This returns 500 photos, so he switches to map view, and looks at where the pictures were taken (GPS has auto-located the photos). Selecting a certain neighborhood, he is able to find a good photo of what he wanted. He also browses through the photos, marks a number of them as “share with friends”, marks all the events in his calendar as public, and an attractive story is automatically created on his blog (with access limited to his friends).

The next morning, Gordon’s body sensor (Figure 2 - middle) tells him he has a fever. His analysis software notes that he has been getting sick after business trips recently, and he forwards the analysis

to his doctor to get advice on how to avoid this. He still feels well enough to go to work, but can't find his hat. The last time he remembers seeing it was after doing his laundry. He accesses the log of his washing machine and finds the last time he used it. He asks for photos taken by his room-mounted cameras in his home just after that time and quickly scans through them until he sees a shot of himself tossing the hat on his bedside table. He looks behind the table and finds the hat.

Figure 2 – Left: wearable video camera from Deja View. Middle: wearable bio-metric sensors by BodyMedia. Right: LG Internet fridge

3.9.2 **Overview**

Everything is becoming smart and networked: objects like refrigerators and pens; places like meeting rooms and living rooms. A/V capture is becoming wearable. Bio-metric sensing is blossoming. The era of abundant storage we are entering makes keeping most of one's life possible. The era of networking promises to allow one to view and manage from any device, any place. The combination of this technology will let us capture most of our lives in a passive way, so that one will no longer need to stop interacting to become the movie or picture taker.

We discussed what one would do with a complete life of digital memories, and looked at the possible applications over the course of a lifetime. We considered reasons why one might not want to keep everything, outlined some research challenges, and also identified the unique leverage that having a complete (across both time and data types) life record brought to PIM problems.

3.9.3 **Description**

Perhaps the first question everyone asks about a completely digitized life is: why? What would you do with it? Some of the obvious answers are:

1. Recall
    a. Find things (such as keys and eyeglasses)
    b. Replay learning and teaching experiences
    c. Review past research and trips to places
    d. Remember names of people and places
    e. Discussions in meetings
2. Share experiences with others
    a. Relive experience of lost loved ones
    b. Grandparents to grandchildren
    c. Revisit a personal experience again

3. Personal reflection and analysis
    a. Understand personal development
    b. Review conflict situations
    c. Find patterns that are common to emotional states
4. Time management
    a. Improved health via medical monitoring

Figure 3 illustrates the space of applications by who controls and uses a person's digital memories. The applications will change over the course of one's life, as will the person using the application. It early life, one's parents and caregivers will use and control one's information. For example, your parents or the babysitter may need access to your dietary and homework information. One's adult life will be a phase of personal control and use. At the end of one's life, caregivers will again take a lead role, needing to look at your medical history or helping schedule appointments in your calendar. After one's life, the executor of your will can access your financial information, while your descendants can learn about their roots from your digital memories.

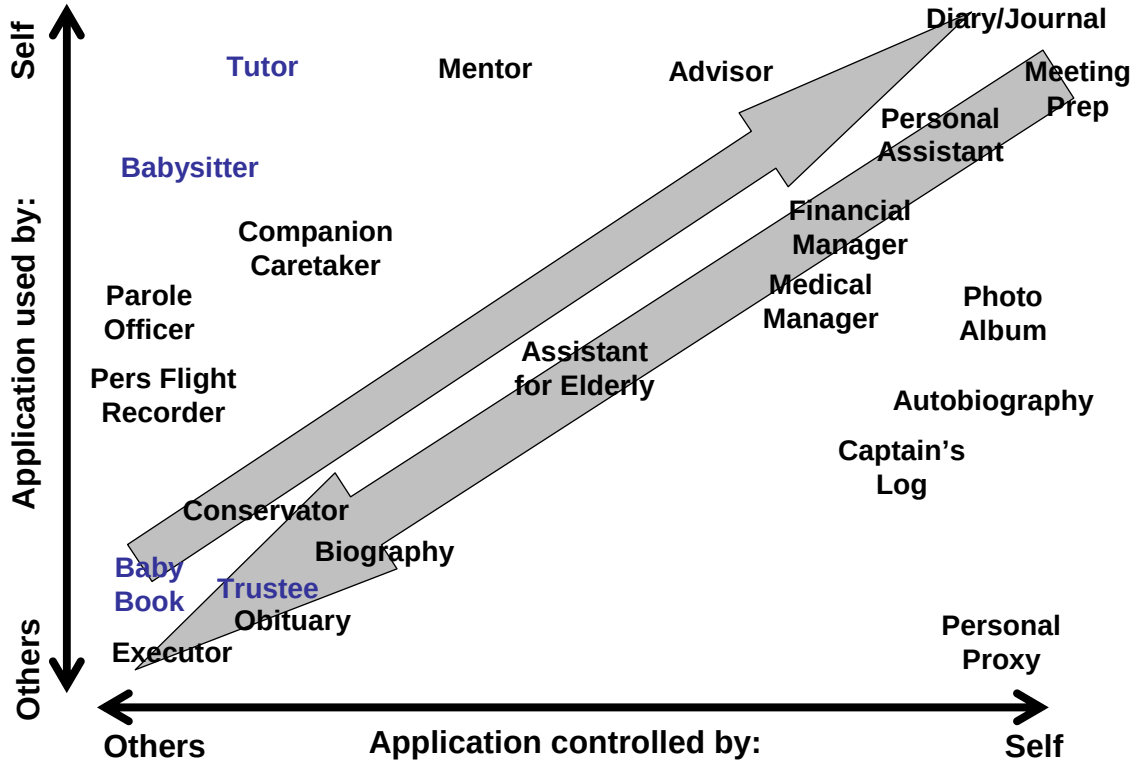

Figure 3

It is fascinating to speculate about inferences that can be derived from digital memories. Poor health might be correlated with certain locations or activities. The onset of poor health may be detected before the user is aware of the condition. Perhaps the system could even identify a pattern of poor choices in dating partners.

There are, however, arguments to be made against keeping everything in one's life. It may be that I don't want an accurate memory of the past in some instances. The memory may be painful. In fact, psychologists indicate that forgetting negative events (or at least having the member fade and become less clear) is critical to recovering from trauma. I may also want to forget for legal reasons: I don't want my digital memories subpoenaed so that they can be forced to testify against me in a way my real memories cannot. Plausible deniability may be lost if I keep a recording of everything. Furthermore,

there are privacy concerns: the more I store, the more might be seen or stolen. Admittedly, this is only a quantitative difference from the privacy concerns I have with my PC today, not qualitative. Nonetheless, it makes a scary situation scarier. Perhaps it is beneficial to receive a good scare now so that we will begin dealing with the issue rather than gradually creeping towards a privacy crisis[5].

There is also a natural concern that clutter could keep one from finding things and that the increased volume would lead to more management work for the user. However, it also seems clear that anything one might have deleted could just be marked to be hidden by default; this would eliminate the clutter and management while still retaining the item for possible use in the future. Furthermore, it is often impossible to predict whether some thing will be needed again in the future. This is why some of us have large filing cabinets full of paper today: not because we will want to access every piece of paper, but because we cannot predict the very few papers that actually will needed again. Keeping as much as possible avoids destroying something that may turn out to be valuable after all.

Fortunately, the scaling of the number and types of data is an advantage as well as a challenge. With the increased scale comes more opportunities to correlate – mostly likely based on time or place, but possibly on any common attribute – and such correlation can clearly be leveraged to help find things and to help users tell stories about their lives. For example, having a personal location record allows one to find the document you edited while in Seattle, and photos can be connected calendar events with the same time value to turn a calendar into a photo-diary.

3.9.4 **Key research challenges**

The primary research challenge for digital lifetime memories is coping with the sheer quantity of material. Summarization, abstraction, and data mining approaches must be investigated to identify "important" items, although what is important to one person is not important to another, and what is important today may not be in the future. Multiple levels of detail and resolution are desirable for all captured media, and especially sensor data.

Making use of the increasing number and types of data sources (primarily from sensors) poses another challenge. The information must be abstracted and displayed in useful and attractive visualizations. It is not clear that all the details of all of one's devices should be part of one's digital memories; perhaps devices will have their own digital memories and the user will only want summaries. For example, my washing machine may have the complete record of all its RPM values sampled every ten seconds, while my digital memory only cares how many loads of laundry I did.

If the question is just whether to go digital or not, security is not an issue. I can lock up my hard drive in exactly the same way that I would have locked up my papers and photos. Furthermore, it is easier to make a perfect copy and store it at another location for disaster survival. However, the convenience of access makes it desirable to attach my digital memories to the Internet – one of the key premises of this breakout is ubiquitous access. Now, instead of one locked door in one place that a burglar may attack to get at my data, I have put a digital door into millions of virtual neighborhoods for every burglar in the world to take a crack at.

Even presuming that systems can keep my private information safe, simply specifying what should be private or public is a challenge: it is critical to get right for privacy, so it must be handled correctly, but without imposing undue burden on the user. Even the binary choice of private vs. public could be onerous, but a really satisfactory solution should have designations of exactly who information gets shared with and when. Different layers of security may be desirable with different types of information having a default layer assignment.

3.10 Beyond Email …

[5] People are generally ignorant of the level of risk to their privacy they face right now, and don't realize how much it has been compromised already.

Steve Whittaker, Jacek Gwizdka, Tom Erickson, Jonathan Grudin, David Levy.

The discussion was organised around debate of what we hoped were controversial statements about email, each of which have implications for its future.

1. Email is the killer app and should the general focus for PIM

The argument is that email is already a multi-functional application. People use it as a file system, communication manager, todo list, contact manager… So let's acknowledge this, and explicitly integrate these other functions directly into email.

Problems:

(a) **Integrating more functions into email would mean a disaster.** Email isn't that efficient at its 'own' functions anyway. People complain about email overload, as well as not being able to find information, contacts or tasks in their email. So to overload email still further with more applications would be to court disaster.
(b) **People prefer a variety of applications**. Empirical studies suggest that people want to use an ecology of applications, where they focus on the 'core competence' of each of a suite of applications. It follows from this that the best strategy might be to encourage users to migrate some of email's usages to other applications, that are specifically designed for the purpose, e.g. using a dedicated contact manager rather than the email address tool. Actually a more nuanced position might be to provide data level integration of these functions and to allow users access to that data via multiple interfaces.
(c) **Population differences**. Other empirical studies suggest that email isn't the killer app for all populations anyway, and that students for example express a strong preference for using Instant Messenger as their key application. So integration around email clearly wouldn't help the student population. This also implies that there may not be a single killer app for all populations. A conclusion might be drawn. If we follow the killer app integration strategy then the integrating app will differ across populations, making the strategy harder to implement, undermining the unifying nature of the killer app, and possibly the whole integrative strategy.

2. Email is completely broken and we need alternate models for managing information delivered through it

The argument here is that people despair of email, because: they get too much information because of spam and careless broadcasting behaviour, they can't find working information relevant to their current tasks, they find it hard to file email information in such a way that they can find it again. Other studies indicate that email traffic and overload is contributing to work-related stress. Furthermore, these problems can only get worse as the number of email messages sent over the last 3 years has increased 8 fold. Although we noted that email is a legacy application which may now be part of the communication cultural (and hence hard to change) we nevertheless discussed several alternatives to email:

(a) Separating transient versus longer-term communications:

IM plus blog, using IM for communication plus the blog for publishing. It remains to be seen how successful this approach is.

(b) Collaborative project centric information management:
the problem here is the classic workflow problem that not enough of email can be organised into projects to make this a useful unifying organisational principle. In other words, not enough tasks can be organised into collaborative projects and too many messages are singletons that have to be processed in isolation.

This last topic led to two related discussions about task inference and management (3) and workflow (4).

3. Task inference and management is the key to improving email

The argument here is that many of the problems that arise in managing email result from the fact that it's hard to access and organise information relating to the same work task. It follows that if we could infer such tasks then we could make email better within either of above approaches.

(a) several of the group participants were skeptical that we will be able to successfully infer tasks. This has been a classic problem in both HCI and psychology for many years now, and not much progress has been made into task inference. Having said this, much progress has been made recently in areas such as machine learning and text processing, both of which may allow email messages to be analysed in promising new ways.
(b) Again, however, one worry with these techniques is that tasks don't account for enough of email's complexity. In other words even if we could successfully identify a large proportion of email tasks this would still leave a large residue of messages to process that are singletons and not part of any task. However some empirical work might be useful here to determine what proportion of people's email concerns sets of messages related to specific tasks, as opposed to unrelated messages.

4. Email is workflow in disguise (similar to task management)

Most work is collaborative. If we own up to this we could incorporate ideas from workflow, including

(a) better tools to track collaborative tasks
(b) lightweight features that would help people to manage collaborative tasks
(c) but the problem is that this approach has been tried multiple times (e.g. Malone et al., Lotus Notes) with little success. One problem here is that workflow doesn't seem to cover a large enough proportion of users' tasks. Again there seem to be too many messages of other kinds.

5. Redressing the balance between senders and recipients

One major problem for email recipients is their lack of control of the volume of messages that they receive. While spam is a major contributing factor, the issue applies also to messages from "legitimate" sources. We talked about several approaches to deal with this:

(a) filtering – using AI techniques to build user profiles that would allow intelligent filtering of messages. These techniques could be used to deal both with spam as with non-spam emails. These techniques are improving, but work is still needed to improve the programming interface to these.
(b) We talked about experimenting with other methods to control spam, one might be to charge people for sending messages (pay-to-send). Another idea we talked about was to use reputation systems (or similar techniques) to identify important senders of email, so that their messages might receive precedence (or at least not be deleted).

(c) We also talked about educating people about email sending habits, but we were somewhat skeptical about whether such methods were likely to succeed.
(d) Another thing that we discussed with respect to dealing with non-spam messages, was changing senders' expectations by reducing their expectations that every message will be read. We talked about how if people became used to recipients use of filters that they might start to have decreased expectations that every message they send might be read – which might in turn modify their sending behaviour.
(e) A final approach might be to try and have senders do more work to provide information about messages (e.g. semi-structured messaging). But again part experience suggests that this approach may not work.

6. Searching will solve the email problem

Argument here is that offered by Google's gmail, that the main problem with email is *finding* messages,

(a) but this ignores the fact that many of the problems with email are in deciding what action to take with new incoming messages, and in tracking the status of undischarged messages. Part of the function of the inbox is to serve as a reminder about undischarged messages and it isn't clear that a search only model can address this.

7. Email needs to incorporate 'pull' type components.

Instead of having all information sent directly to users we need to experiment with techniques whereby information that is not directed at a single individual is published at a public location rather than being sent to an individual.

(a) now several tools, blogs, combine blogs with documents stores to address some of the problems of version tracking
(b) problem of deciding what should be published rather than pushed to people
(c) if information is published do we have some form of alerting to tell people where that information is located (otherwise they may not know of its existence)
(d) Problem with alerting is that this may be almost as distracting as the original message
(e) Also users have to know *where* information will be published. If people don't know this then they may not be able to find the information. Even though email may be overloaded at least people know that the information they require is located in their system.

4 Conclusion

During workshop discussions, several PIM challenges & issues emerged:

Information is fragmented; so too, is the study of PIM. The information required to complete a task – planning a trip, for example – is frequently scattered across physical locations (home, work) and devices (a PDA, a laptop, etc.). Information is often stored in different organizations according to its form. A person may maintain one or more separate organizations for each of the following forms: Paper, email, electronic documents, web references (bookmarks, favorites), calendar entries… and the list of forms continues to grow. (For example, Microsoft's new *OneNote* application provides a tabbed method for organizing notes separate from the file hierarchy, email, calendar, etc.) Gathering the information needed to complete a task can then be a major chore in its own right. With multiple locations, devices and information organizations the chances for confusion and inconsistency increase as well (so that, for example, a person ends up looking in all the wrong places for a desired piece of information).

The study of PIM itself is often fragmented in similar ways. Many excellent studies focus uses of and possible improvements to email; other studies similarly focus on the use of the Web. Of course no single study can address PIM in its entirety. But in defining a study along the lines set by existing applications and information forms, we may miss important opportunities for information integration.

How to protect the privacy and security of persona information? The more complete our personal information, the more completely someone else can assume our identity. New tools of PIM – especially those aimed at information capture – must be accompanied by new levels of information security. How can we audit the information about us held by others? For example, if we could determine everywhere our Social Security Number appears, we could weed out inappropriate use (for example, clerical errors or identity theft).

Where do the bits and pieces go? Calendars contain appointment information; address books contain contact information. But many items of information seem to fall through the cracks between existing tools of PIM. Example: "A good hotel to stay at in Seattle is the Watertown".

Who owns the information in the workplace? Suppose, for example, that a PIM system is able to capture an employee's experiences and the knowledge she gains on her job. Who gets this information if she decides to leave for another company?

How can an employee's knowledge of the information space be captured for later use? For example, Boeing service engineers are specialists (such as in avionics or hydraulics) who answer queries about aircraft maintenance and repair for field engineers. To answer a query, a service engineer may track down a wide range of information to formulate a response: engineering documents, airframe history and modifications, maintenance procedures, FAA regulations, minimum equipment lists. Currently every query and response is captured, to provide an aid in answering similar questions in the future, but the set of information items consulted is not recorded, which might serve as a guide to a new service engineer.

How do we know what is working and what isn't? Evaluation of new PIM tools and techniques is very difficult for a number of reasons: a.) the tool/technique may help with one aspect of PIM but hinder others. It is necessary to evaluate the overall effect of a tool/technique on an individual's ability to manage information. b.) PIM tools/techniques cannot be easily evaluated in a laboratory setting. Management of information occurs against a backdrop of other information and everyday tasks. A synthetic benchmark or common information collection can't very well play the role of an arbitrary subject's personal information space. c.) People adapt and their needs change. An accurate picture of a tool or technique's utility emerges only over an extended period of evaluation

How can we keep PIM concerns from "falling through the cracks" as new tools and technologies emerge? Workshop participants repeatedly expressed a concern that larger issues concerning how we manage our information across tools and over time were overlooked and misunderstood in the rush to "ship". Several participants felt that technology seems to have lost its luster as "the answer" to our problems. Participants recalled earlier hopes for applications and new devices that remain unfulfilled.

4.1 Recommendations

The PIM05 workshop produced the following recommendations for the National Science Foundation:

1. Encourage multi-disciplinary approaches. Expertise relevant to PIM comes from range of academic disciplines including cognitive psychology, sociology and social psychology, data management, information retrieval, human-computer interaction and also from domain experts (in medical informatics, for example). Participants of a research project should ideally be able represent two or more of these disciplines.
2. Research into promising PIM tools and technologies should be balanced by empirically grounded studies aimed at acquiring a better understanding of underlying problems of PIM. Several important discussions in the literature ("Will filing go away?", "Is it worth it for an individual to archive information?") remain stuck at a level of yes-or-no questions. Data on actual practices of PIM in different situations can help to parameterize key questions ("What is needed for filing to go away or get easier?", "Under what circumstances is worth it for the individual to archive information?" "What kinds of information should we be sure to capture?" "What aspects of PIM are especially problematic for the individual?"). Answers can help to guide tool-building efforts. Observational studies of people in actual situations of PIM, though expensive to conduct, are likely to be especially useful. Furthermore,

   a. Shorter-term, “point-in-time” observations should be balanced with longitudinal studies in which patterns in an individual’s practices of PIM are mapped over a period of weeks or even months.

   b. Broadly-based studies of people in a diversity of information-intense activities should be balanced with deeper look a professionals in selected occupations. Specifically noted as worthy of study were the PIM challenges that physicians and other clinicians face. The PIM needs of intelligence analysts – especially in the area of homeland security – are another obvious and important area of study.

3. Support the development of methodologies, frameworks and benchmarks for the evaluation of PIM tools and techniques. The workshop recognized that the evaluation of PIM tools and techniques is very difficult for a number of reasons: a.) the tool/technique may help with one aspect of PIM but hinder others. It is necessary to evaluate the overall effect of a tool/technique on an individual’s ability to manage information. b.) PIM tools/techniques can not be easily evaluated in a laboratory setting. Management of information occurs against a backdrop of other information and everyday tasks. c.) People adapt and their needs change. An accurate picture of a tool or technique’s utility emerges only over an extended period of evaluation.

4. It is important that at least some of the research take broad view of PIM. As noted above, research into PIM, like personal information itself, is too frequently and artificially fragmented along the lines of specific applications such as email, electronic file management or web browsing. Progress in PIM will require integrative approaches that help people to manage their information in a consolidated way according the tasks they must perform and across the various types of information that must be managed – including audio, video as well as text.

5. Get organizations, including both government and for-profit, involved. As noted above, improvements in PIM can benefit the organizational bottom-line in several ways – through increased employee productivity, better collaboration among team members and, longer range, through better management of employee expertise.