

Personal Information Management Group Report²

William Jones

The Information School
University of Washington
Seattle, WA 98195

williamj@u.washington.edu

David Maier

Dept. of Computer Science and Engineering
OGI School of Science & Engineering
Oregon Health & Science University

Beaverton, OR 97006

maier@cse.ogi.edu

1. Executive Summary

Personal information management or PIM is attracting increasing attention as an area of study. In the ideal, we have the right information at the right time, in the right place, in the right form, and of sufficient completeness and quality to perform the task at hand. Further, we can manipulate that information with ease, to organize, excerpt, annotate, group and link it to accomplish the task. Tools and technologies help so that we spend less time with burdensome and error prone activities of information management (such as filing) and more time making creative, intelligent use of the information at hand in order to get things done. The payoffs for advances in PIM are large and varied. 1.) 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. 2.) 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 3.) improvements in education programs of information literacy and 4.) better support for our aging workforce and population.

PIM creates special opportunities for the application of both Information retrieval and data management. For example, a large, digital store of persona information can be used to customize a query, the results it returns and their display. This store can also be mined in order to render important patterns and connections into more explicit, useable, and updateable forms. Techniques of automated inferencing, clustering and classification all have potentially important applications in support of PIM.

But PIM also faces significant challenges. The information we need to complete a task is often scattered across locations and devices. Information is further fragmented across separate organizational schemes – for paper, electronic documents, email, web references, etc. The opportunities to support PIM increase as increasing amounts of personal information are captured in digital form. But the capture of personal information raises important issues of security and privacy. The capture of personal information in the workplace raises additional issues of ownership. The field of PIM must also devise ways to evaluate candidate tools and techniques.

The discussion group had several recommendations for the National Science Foundation: 1.) Research into PIM should involve expertise from a variety of academic disciplines including cognitive psychology, sociology and social psychology, data management, information retrieval, and human-computer interaction. 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. Some of this research should focus deeply on the needs of selected professionals (doctors, for example). Though expensive to conduct, some research needs to extend over a period of time (weeks or even months) in order to better understand how PIM changes for an individual with time and to discern long-term patterns of use. 3.) It is important to support the development of methodologies, frameworks and benchmarks for the evaluation of PIM tools and techniques. 4.) It is important that at least some of the research take broader, integrative view of PIM that looks across and beyond existing applications and information forms (e.g., for email, e-documents, web references, “notes”, etc.) to images, audio, video and even sensor observations. 5.) It is important to get various organizations – private, non-profit, governmental, etc. – involved in the support of PIM research. 6.) One or more workshops specifically targeting

¹ The workshop was held Sept. 14-16, 2003 in Seattle, WA.

² With contributions from Harry Bruce, Ramani Duraiswami, Ben Bederson, Gregory H. Wakefield and others.

PIM could significantly help to establish current imperatives and priorities of PIM as a field of study and help to bridge across disciplines.

2. Background and Motivation

The discussion group began with a consideration of what Personal Information Management (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.

In the literature, PIM is more often expressed as a future goal state or ideal rather than a current reality. Good PIM means that we have the right information, at the right time, in the right place, in the right form, and with sufficient completeness and quality to complete the task at hand. Further, we can manipulate that information with ease, to organize, excerpt, annotate, group and link it to accomplish the task. Tools and technologies help so that we spend less time with burdensome and error-prone activities of information management (such as filing) and more time making creative, intelligent use of the information at hand in order to get things done. In the ideal, we have more time to be human beings.

The current reality of PIM is otherwise. The information required to complete a task is often scattered across many locations and in many forms. It is buried somewhere in email, in the paper documents of a stack, or in a web page visited last week. Information we need for a task at work may be at home and vice versa. Considerable time is spent creating and re-creating folder structures in which to store information (one for electronic documents, another for paper documents, another for email and still others for notes, contacts, web references, etc.). Time may be spent organizing information that is never used. And filing is sometimes counter-productive. We may forget where we file the information or we may forget to use information until after the task to which it applies is completed. Complaints of information overload are widespread and many of us frequently have the feeling that our information is managing us rather than the other way around.

The discussion group noted that information can be “personal” for different of reasons: a.) It might be “stuff I’ve seen” – encountered at some point in the past. b.) It might be “stuff I should be seeing” – of use to a project or area of interest. c.) It might be “stuff about me” – information, for example, concerning credit history, web usage, medical history, etc. That is, PIM is understood both as personalized management of information and management of personal information.

Independently of these variations, personal information can be d.) “in here” as part of a personal collection of information and, at least nominally, under a person’s control or the information can be e.) “out there” and managed by others, whether on the Web, in databases, or, more conventionally, in paper archives. Information about us is gathered legally through our contact with doctors, HMOs, insurance companies, the IRS, banks, credit card companies, etc. Information about us is also gathered in semi-legal and illegal ways as well – through the transmission of “cookies”, for example, or through “spyware”. Our awareness of what personal information exists in the hands of others is low.

The NSF has taken a lead in funding investigations into the issues of privacy and security, policy and law that arise with respect to the maintenance and management of a person’s information by others. In the long run, technology may also help – for example, by enabling individuals to more closely monitor the access to their information and possibly even to trace this information as it moves from place to place.

The discussion group turned its attention to the more direct applications of technology in meeting the classic challenge of PIM: How to insure that information is available, at the right time, in the right place and in the right form, to meet our information needs? Research on this classic problem of PIM is motivated by the enormous potential for payback at several levels:

1. **For each of us, as individuals**, better PIM translates into a more effective use of our time, energy and attention. As the many testimonials from popular books on PIM will attest, bad PIM is a drain on our energies and a leading reason for feelings of being overwhelmed and out of control. Good PIM directly affects our quality of life. The observation was made that, while the supply of information is growing at apparently exponential rates, we are “wired” with certain limitations as human beings (e.g., in short-term or working memory, in our ability to multi-task, in our ability to remember things accurately over time). PIM holds the

promise of helping us to bridge the gap between an ever growing supply of information and information processing abilities.

2. **Within organizations** there is increasing awareness that PIM is a significant determiner of personal and team productivity. Bad PIM costs time and money. One survey, for example, finds that the average manager in the United States spends three hours a week looking for documents that have been lost or misfiled. Good PIM means that important information is available when decisions must be made. Good PIM can help organizations in at least two other ways:
 - a. Good PIM is a necessary step towards better management and leverage of employee expertise.
 - b. Good PIM can also mean a better management of information within teams and larger groupings of an organization. Some benefits are clear. For example, a group may depend upon information maintained by an individual. Other benefits are less obvious. For example, good PIM may help to make explicit key differences in the way different team members view information. There is then an opportunity to deal with these differences directly.
3. **Educational programs in information literacy** can improve with the study of PIM. Progress in PIM is made not only through new tools and technologies but also with new teachable techniques of information management. For example, people may need to learn techniques of sampling if they are to cope with the tremendous amounts of information returned by web-based search services in response to most queries. Information returned is of uneven quality and much of it is redundant. The challenge then is not to find the (complete) right answer but rather to find a reasonable sampling of the information available.
4. **For an aging workforce and an aging population**, better PIM may increase the chances that our ‘mental lifespan’ matches our physical lifespan. It is well-established, for example, that the span of short-term or working memory decreases steadily with age even as knowledge gained from experience increases. Better PIM can reduce the need to keep several things in mind at once, thus reducing our reliance on working memory. PIM can help compensate for forgetfulness or difficulty in committing new people or experiences to memory.

3. Opportunities

Given this year’s theme for the IDM workshop, “Information Retrieval and Databases: Synergies and Syntheses”, a discussion group on PIM is especially timely. An emphasis on PIM creates special opportunities for the application of both Information retrieval and data management. For example, we can assume that more and more information will be kept by, for and about the user in digital form. It is, therefore, perfectly appropriate to explore the implications that such information might have for the actions of search. Queries, the results returned and their manner of display can potentially reflect a lifetime of information concerning user, her current projects and enduring areas of interest, her preferences and her previous experiences. Searches can, of course, be ongoing with results integrated more seamlessly into structures of information organization (e.g., on-line calendars, email, folder hierarchies) with which the user is already familiar.

Techniques of data mining and data structuring can also be applied to a store of personal information in order to render important patterns and connections into more explicit, useable, updateable forms. The inbox, for example, contains a wealth of information concerning the people, projects and areas of interest in our lives. It was also noted, for example, that information regarding an individual (email address or phone number, for example) is often scattered across a range of devices (cell phones, PDAs, different computers, different applications such as an address book or a calendar on the same computer, different physical locations, etc.) in separate and often inconsistent versions. Wouldn’t it be nice if all information regarding a person (company, project, topic of interest, etc.) could be accessed easily in a single effort (regardless of device or location)? Wouldn’t it be nice if data were normalized so that an email address, for example, needed to be updated only once? The group also noted that personal repositories of information such as an email inbox or a history of web sites visited represent opportunities for data mining. An analysis might reveal, or help make more explicit, key people, projects, dependencies, etc. that are deeply embedded and widely scattered in the repository.

We can move a step further into the realm of knowledge representation and artificial intelligence. A “smart” PIM system might reason from a person’s information store, including its schedule of events and tasks to complete, in order to make assumptions regarding the given and the new, the relevant and irrelevant in incoming information. Attention, perhaps a person’s more precious resource, could then be directed first to those parts of the incoming information that are most “informative” – new to the person and also relevant. Such a system might, for example,

direct the attention of a doctor to revised practice guidelines for treatment of a condition of a patient under her care, or alert her to an outbreak of a new or unusual disease (such as SARS or anthrax).

More possibilities arise if the patterns of information access recorded in one person's information store can be shared with colleagues, possibly as overlays to the information they must process. An example was given of a physician who, while paging through a patient's (rather thick) medical record, noted that the patient had seen a cardiologist recently, and remarked, "I wish I knew which items the cardiologist looked at". Many of us may have a similar feeling. We want to know what our respected colleagues, from the vantage point of their own special expertise, are looking at and think important. The attention of experts is a scarce resource, and reusing expert attention by capturing their information selections can have high value. The more complete the digital record of a person's interactions with information becomes, the more possible these examples become. We then realize Vannevar Bush's notion shareable "associative trails" as overlays to the information.

PIM also creates opportunities for the application of clustering and automated classification techniques. Clustering techniques might be used, for example, to suggest new ways of organizing a person's store of information. A person's organization(s) of information, regardless of how it is derived, is a special kind of taxonomy. Its concepts (people, places, projects) and their organization are at least partly a reflection of a person's world view. An ordinary folder/subfolder hierarchy, for example, is a kind of personal classification scheme. What can a PIM system learn from category exemplars (i.e. the documents, email messages, web references, etc. that are stored within the folders)? And how can this knowledge, once induced, be applied to facilitate the classification (filing) of new information?

Consider a "Personal Health Album" for an individual. Much of the data and documents relevant to a person's health exist (somewhere) in digital form: medical records, exercise plans, appointment schedules, drug information, vaccination records, insurance forms, and family medical history. However, that information is for the most part not organized from the point of view of the consumer, but rather from the perspective of the medical provider (or, more likely, multiple providers). With a Personal Health Album, a person could take all this health information and cut it up and paste it on pages in a way that made sense from his perspective. There might be a page for all current medications, with prescriptions, dosages, refills remaining, insurance payments, and relevant excerpts from the information sheets that come with drugs. (For example, he might not include the "warnings for pregnant women" section, but may include "interactions with alcohol"). He can create another page with all information related to a knee injury, such as providers seen (urgency care paramedic, radiologist, orthopedist, physical therapist), diagnoses, MRI images, care and exercise regimen, and future appointments.

4. Challenges

The discussion group considered the following challenges of PIM:

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 do we capture information from our lives away from the computer (and other electronic devices)?

Some of us felt that we didn't have much of a life anymore that was separate from our "devices" (computer, mobile phone, pda, etc.) but this is a problem for another breakout session ☺. One member of the group gave the example of his voice lessons, which he routinely recorded. It would be very useful to review sessions especially to

look at the times when the instructor said “good”. Someone expressed a desire to capture those “great ideas or insights” that are often gone in the time it takes to find a paper and pencil (or some other recording device).

How do we keep others from capturing and disseminating our 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 Madison”.

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 make more effective use of existing tools and technology? The group expressed a feeling that many of the pieces needed to improve PIM are already available. The challenge is to assemble these pieces in a way that works for people. The group recognized the “Stuff I’ve Seen” (SIS) work described in one of the plenary talks by Susan Dumais as very promising example of this “assemble from pieces” approach.

5. Recommendations

The discussion on PIM resulted in 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 group 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.
6. One or more workshops specifically targeting PIM could significantly help to establish current imperatives and priorities of PIM as a field of study. Participants should represent a range of academic and domain expertise and should be drawn from industry, academia and also from government agencies.