PIMs Workshop Report: Measurement and Evaluation

Diane Kelly (Facilitator, Report author)
School of Information and Library Science
University of North Carolina at Chapel Hill
www.sils.unc.edu/~dianek

Ben Bederson
Human-Computer Interaction Lab
University of Maryland
www.cs.umd.edu/~bederson

Mary Czerwinski
Microsoft Research
http://research.microsoft.com/users/marycz

Jim Gemmell
Microsoft Research
http://research.microsoft.com/~jgemmell/default.htm

Wanda Pratt
The Information School & Biomedical and Health Informatics
University of Washington
http://www.ischool.washington.edu/wpratt/

Meredith Skeels (Scribe)
Biomedical and Health Informatics
University of Washington

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.

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.

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. 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.

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.

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.

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.

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.

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.

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.

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.

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.