Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2008-007 February 10, 2008

Wicked Problems and Gnarly Results: Reflecting on Design and Evaluation Methods for Idiosyncratic Personal Information Management Tasks

Michael Bernstein, Max Van Kleek, Deepali Khushraj, Rajeev Nayak, Curtis Liu, mc schraefel, and David R. Karger

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Wicked Problems and Gnarly Results: Reflecting on Design and Evaluation Methods for Idiosyncratic Personal Information Management Tasks

Michael Bernstein[1], Max Van Kleek[1], Deepali Khushraj[3], Rajeev Nayak[1], Curtis Liu[1],
mc schraefel[2], David R. Karger[1]

[1]MIT CSAIL
32 Vassar Street
Cambridge MA
{msbernst, emax, karger} @
csail.mit.edu

[2]Electronics and Computer
Science
University of Southampton
Southampton, UK, S017 1BJ
mc+chi@ecs.soton.co.uk

[3]Nokia Research Center
Cambridge
3 Cambridge Center
Cambridge MA
deepali.khushraj@nokia.com

ABSTRACT

This paper is a case study of an artifact design and evaluation process; it is a reflection on how right thinking about design methods may at times result in sub-optimal results. Our goal has been to assess our decision making process throughout the design and evaluation stages for a software prototype in order to consider where design methodology may need to be tuned to be more sensitive to the domain of practice, in this case software evaluation in personal information management. In particular, we reflect on design methods around (1) scale of prototype, (2) prototyping and design process, (3) study design, and (4) study population.

Author Keywords

Case study, user-centered design, personal information management

ACM Classification Keywords

H5.2. User Interfaces: Evaluation/Methodology

INTRODUCTION

This paper is a case study of the user-centered design process in action -- how a design and research team moved from a seemingly strong grounding in prototypes and expert feedback to a user study that returned unexpected results. After months of in-depth research of related work, preliminary ethnographic studies [10], initial prototyping [30] and deeper observational studies of work in practice [9], our research team arrived at a set of hypotheses we determined needed to be tested in order to evaluate the system concept. We carefully reviewed the hypotheses against all our primary and secondary data to assure ourselves that we would be testing (a) the minimal set of functions and (b) the minimal set of hypotheses to reach conclusions about our approach, with our goal being a submission of our findings to a prestigious human-computer interaction conference. We had determined as well that due to the nature of the system, we needed to test this iteration of the prototype both longitudinally and with participants in the wild. We checked sources to look for comparable studies to see what the usual deployment times were for such studies.

With our approach feeling well grounded, we designed and implemented the system under deadline pressure, succeeding in crafting the research prototype in time for the study. We met with our participants, trained them to use the system, and kept in contact with them during the week they used the system as part of their regular practice. Participants' feedback, however, was unexpected: it focused mainly on users' pre-established practices (which we had already investigated), giving us little feedback about the system and our actual hypotheses. Our initial reaction to the study results was that we had failed somewhere – that we had done something *obviously* wrong to have gotten responses so strongly questioning the basic design points of our system, rather than the research hypotheses we intended to test.

In reflecting on our design, development and study process against the backdrop of user-centered design methods, however, we found no singularly impressive misstep that set us off-course. Indeed, even now, after reflecting on the process, while some approaches for future steps have come out of the process, it is not entirely clear that those next steps have the backing of current methodology to support them, or will guarantee the desired outcomes.

In this paper, we present our design process as a case study framed against a variety of usability and design methods in order to investigate where our adherence to some methodologies might have been too weak, and in other cases may have been too strong. Our goal is to identify possible gaps or research opportunities for shaping design methodology, and to see as well whether the domain under investigation might itself not require a particular kind of design approach that differs from user centered design of other application types. In the following sections, therefore, we review what we were attempting to build and how we rationalized our investigative approach. We then look at three points in particular in the study roll out process that our reflection on process suggests might be the break points in the process: (a) scale of prototype (b) choice of participants (c) management of participants in a longitudinal study. We interrogate these points against known design methods. We conclude with a consideration of how to move forward, and reflect on implications of our experience for design practice.

DESIGNING A PERSONAL INFORMATION MANAGEMENT TOOL

Our goal was to iterate upon the design of a tool for managing what we call *information scraps* [9, 10]. Based on our own primary studies [9], prototyping [10, 30], and study of related research, we came to the conclusion that we had identified one of Rittel's "wicked" design problems [25] and that a highly iterative, participant-informed process would be necessary to develop a suitable prototype to let us test our hypotheses.

To contextualize our discussion of the prototypes and our hypotheses, we first review the problem space.

Problem Space

Personal information management research has gone to great lengths to assist us in organizing our messy lives. Yet a subset of that information has stubbornly resisted organization: this content lies instead scribbled on Post-it notes, scrawled on corners of sheets of paper, buried inside the bodies of e-mail messages sent to ourselves, and piped into overgrown text files abandoned on our workstations' desktops or in "misc" folders. The scattered data contains our great (and not-so-great) ideas, our sketches, notes, reminders, driving directions, and our even poetry. We call such personal information, *information scraps*, and seek to contribute to the space of earlier investigations into similar phenomena, including notes and to-dos (*e.g.*, [7, 11, 17, 21]).

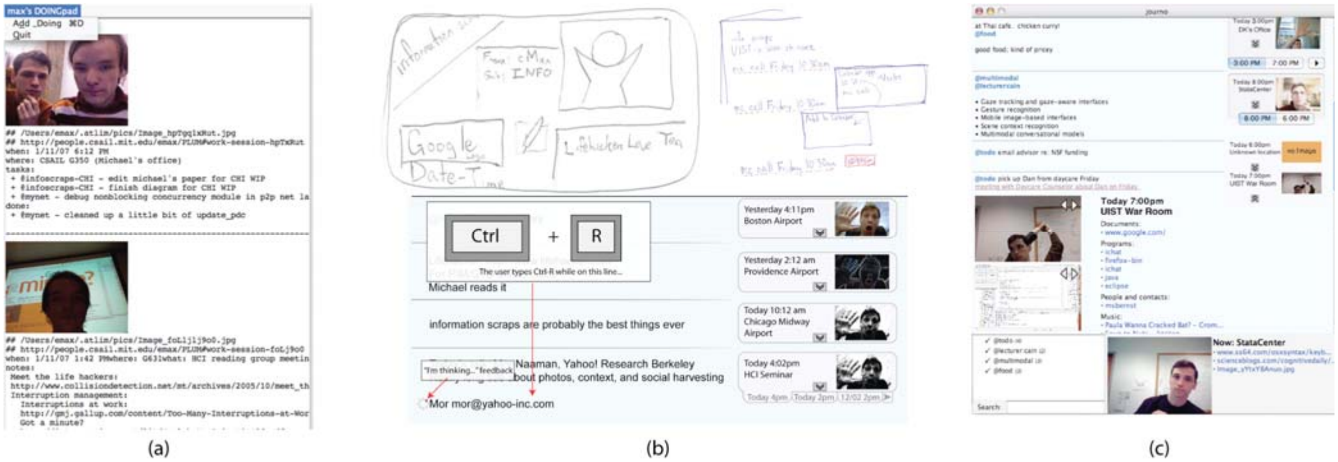
Given their ubiquity, the management of information scraps has proven a difficult challenge. In addition to the numerous structured PIM (i.e., calendaring, task management and email) tool suites that many people use each day (such as MS Outlook), some have taken attempted to assume physical metaphors such as post-it notes [1, 5], or spiral notebooks [2-4] in an attempt to be more suitable for the management of these information scraps. Some users have opted for home-grown information scrap solutions fashioned out shell scripts, to turn plain text files into their own personal information management tools (cite: lifehacker.com, todo.txt.com). Yet despite the large and varied set of digital applications emerging to support this need, no single tool has, as yet, come to satisfy all users' needs, as evidenced by the tendency for people to use a haphazard combination of these and traditional paper-based tools. Our work set out to understand this space, and to improve upon existing approaches.

Process

Early Ideation and Design Space Exploration

Having identified a problem space of interest, we began with a design exercise: if you had a magic text file that could do whatever you wanted, what would you do with it? Our team of four researchers spent a week interacting with this "fake computer." We were interested in considering: what kinds of creative uses could we come up with for this tool if we opened our minds?

We observed a number of interesting characteristics in our logs, for example deliberate ambiguities such as “do ____



```
## /Users/emax/.atlim/pics/Image_hpTgqlxRut.jpg
## http://people.csail.mit.edu/emax/PLUM#work-session-hpTxRut
when: 1/11/07 6:12 PM
where: CSAIL G350 (Michael's office)
tasks:
 + #infoscraps-CHI - edit michael's paper for CHI WIP
 + #infoscraps-CHI - finish diagram for CHI WIP
 + #mynet - debug nonblocking concurrency module in p2p net la
done:
 + #mynet - cleaned up a little bit of update_pdc
------------------------------------------------------------
## /Users/emax/.atlim/pics/Image_foLljlj9o0.jpg
## http://people.csail.mit.edu/emax/PLUM#work-session-foLj9o0
when: 1/11/07 1:42 PMwhere: G631what: HCI reading group meetin
notes:
 Meet the life hackers:
 http://www.collisiondetection.net/mt/archives/2005/10/meet_th
 Interruption management:
   Interruptions at work:
   http://gmj.gallup.com/content/Too-Many-Interruptions-at-Wor
   Got a minute?
```

Figure 1. Evolution of Jourknow prototype: (a) the original DOINGpad prototype, (b) sketches and storyboards for our functional prototype, and (c) the final version.

stuff" or "remind me" notes without any mention of when the reminder should actually occur, and the use of commands such as "open cal." Structure ranged from very orderly notes to almost unparseable text. Verbosity also varied between clearly explicated sentences and very condensed text, even within the same log. Two researchers explicitly recorded contextual information like date, time and location into the text file, and a third, reflecting on his failing memory for the note, remarked that he wished he had done so as well.

Based on our experiences, we built a first prototype system called the DOINGpad (Figure 1a), so named because it captured what the user was doing whenever he or she recorded a note. The DOINGpad was intended as a functional *sketch* [13] intended to explore an idea space -- implemented in four hours, we built it to elicit feedback amongst the design team as we used it. DOINGpad recorded the following whenever the user begins to write a note: current date and time, friendly location name (from the wireless access point; e.g., "max's office"), a webcam photo of the user and his/her surroundings, and a Uniform Resource Identifier (URI) which could be linked to other concurrent system activity such as window switches and music being played. Below these system-generated fields is a free text area for recording the note itself.

Involving Related Work, Functional Prototyping

Having used DOINGpad ourselves to reflect on our approach, we began to iterate upon our ideas. Our explorations spanned several research domains; from PIM research we took note of the inherent tension between a need for lightweight entry and a desire for structured representation later [7, 11, 19]. Studies of remembrance habits then informed us of the various mechanisms our users might utilize to re-find information, such as relevant people and situations [15] or pictures [28]. Here we drew on systems such as ChittyChatty [19], and Stuff I've Seen [16] for design inspiration.

Given this variety of research recommendations, we set out to incorporate them into our tool to see if their insights would positively impact our own work. Through design iteration (Figure 1b), we developed the first version of Jourknow, a journal that "knows." Jourknow represented our first foray exporting our own ideas into the functional prototype space for feedback. Its main design points were automatic context capture and association with notes in support of re-finding (e.g., "it was that note I took down when I was at Starbucks") and lightweight structured expression parsing, which we called our *Pidgin*. We employed first-use studies and design critiques in order to get first-contact feedback on our prototype. The prototype was, however, still too slow and too brittle to be used on a regular basis.

Expert Feedback

With the Jourknow prototype demonstrable but not yet stable or polished, the design and research team decided the next appropriate step would be to put the Jourknow interface to expert critique. We headlined information scraps and Jourknow as a work-in-progress poster at CHI 2007 to gain feedback from the attendees [10]. We received a much more positive than anticipated response (first place award, people's choice), much positive feedback, and many requested features.

At this point we also received expert reviewer feedback on Jourknow, and acceptance of the prototype into a top-tier computer science and HCI conference [30]. Reviews indicated support for our direction but a need to test our ideas on real users:

- "The authors have implemented a reasonably complex system to try to address this well-motivated problem. [...] Since it was informally evaluated with CS students in a lab, how can we know if this is even reasonably usable for non-techies?"
- "There is a need for longitudinal testing to establish how such a system would fit in with people's working practices: who does such a system actually suit, and why?"
- "I agree with the other reviewers that this paper describes a cool system. Certainly I want to use something like this. [...] but, I'm also not sure what we learn from this work without evaluation."

Needfinding and Ethnography

Before incorporating the expert feedback into our prototype, we decided first to hone our knowledge of the domain of information scraps. To this point we had based much of our research on existing literature informing information scrap management (*e.g.*, [7, 14, 21]). However, we found that the literature left unanswered questions: what kind of data is kept in information scraps? What kinds of tools are generally used? What do they look like? What factors affect their creation and use?

Thus, in order to more fully understand the makeup, contents, and needs of information scraps, we performed our own investigation [9]. Our study consisted of semi-structured interviews and artifact examinations of participants' physical and digital information scraps across physical and digital tools. We enrolled 27 participants from five organizations: including local technology firms, and an academic research lab. We interviewed participants about their information scrap habits and recorded examples of information scraps we encountered. Through our study, we uncovered 533 information scraps across the 26 participants (we were unable to perform artifact collection from one participant), and coding each scrap for location, contents, and encoding (text, picture, drawing).

Our results pointed to several new avenues for information scrap solutions:

Large numbers of uncommon items. While participants captured many common PIM types, almost one fifth constituted data types that we observed infrequently across the entire study: for example, fantasy football lineups, words to spell-check, salary calculations, guitar chords, and frequent flier information.

Physical media used for mobility. The most popular physical tools were ones that supported mobility. Paper notebooks captured 37.2% of the physical (non-digital) information scraps we indexed; post-it notes accounted for another 23.7%. Several participants remarked that they used these tools especially because they were portable and more socially acceptable in face-to-face meetings.

Information Scrap Roles. We identified five major roles that information scraps played in our participants' lives: temporary storage, archiving, work-in-progress, reminding, and storage of unusual data types.

Desired Affordances. We synthesized the following design needs that support for information scraps will require: lightweight entry, freeform contents, cognitive support, visibility and proactive reminding, and mobility.

Scoping and Research Specification

At the conclusion of our study, we reflected upon lessons learned and how we might apply our new knowledge to Jourknow. During a two-day caucus, the researchers attempted to scope the project to areas of interest in need of evaluation. We grounded our hypotheses firmly in our own work as well as related research -- each hypothesis needed to be justified by observations from our ethnographic work. For each feature, we examined whether leaving it out would significantly harm the overall effectiveness of the system.

We began with our two hypotheses from the previous prototype: context capture and Pidgin structured language input. An object of discussion was: should we leave the work at those two hypotheses for evaluation, or add something new? We foresaw that users who did not always carry laptops may see limited use to the system, just as users of existing digital tools in our study found mobility a major inhibitor. Furthermore, our participants reported that their tools were often rendered useless when they were not accessible when a note was needed, for instance when away from their desks, driving to work, or at home. Thus, we hypothesized that supporting mobile note-taking might greatly improve the overall experience and usefulness of our system. Thus, we decided to focus our prototype on the following three improvements to existing information scrap practice: context capture, structured capture (pidgin), and mobility.

Jourknow Client Redesign

At this point the research team took the opportunity to redesign the client based on knowledge gained from our previous iterations. We began by generating a large number of basic interface approaches for information scrap management, then built paper prototypes [24] (Figure 2) to investigate the most promising directions: an inbox metaphor, a notebook metaphor, and a search-only interface. We recruited participants from the lab to interact with the paper prototype, which had already been populated with notes; we found that the notebook metaphor afforded a level of spatial memory that participants generally preferred. However, the list interface also seemed to have merits: physical resemblance to a word processor, a logical place to start capturing (*i.e.*, at the end) and an easy metaphor for supporting both automatic and manual arrangement (*i.e.*, sorting). Thus, we brainstormed and designed the remainder of the interface, relying heavily on existing interface paradigms in faceted retrieval (*e.g.*, [31]) to reduce risk. Over a period of the coming weeks, we continued to refine of the design, focusing on Pidgin and context facet panel.

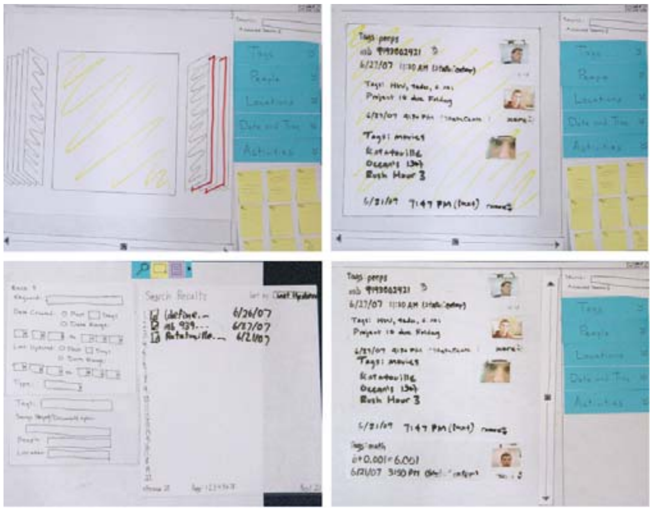
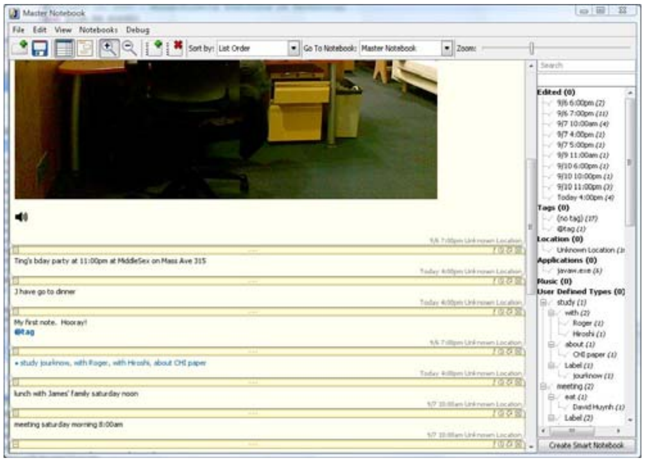


Figure 2. Paper prototypes of the revised Jourknow interface, exploring notebook, list and search approaches.

Prototype Development

A team of four researchers tasked themselves with implementing this new version of Jourknow over an approximately ten-week period. The first four weeks were concentrated on implementing the general client, including the dashboard mechanism, reminders, the basic user interface in a notebook metaphor, and internal logic and representation. Much of the code from the original Jourknow prototype was rewritten to support the new design. Throughout the design and development process, the research team held weekly design reviews with a larger group of students and researchers to get feedback on progress and design decisions.



9/6 7:00pm Unknown Location

Ting's bday party at 11:00pm at MiddleSex on Mass Ave 315

Today 4:00pm Unknown Location

I have go to dinner

Today 4:00pm Unknown Location

My first note. Hooray!
@tag

9/6 7:00pm Unknown Location

• study jourknow, with Roger, with Hiroshi, about CHI paper

Today 4:00pm Unknown Location

lunch with James' family saturday noon

9/7 10:00am Unknown Location

meeting saturday morning 8:00am

9/7 10:00am Unknown Location

Figure 3. The final Jourknow interface.

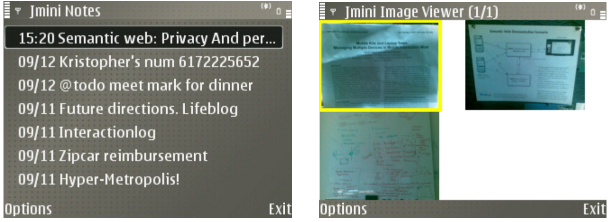
As described above, at the end of the fourth week the client entered a design review and came out with a revised specification. From this point on, our focus was in completing the prototype in time for the summative evaluation to come. Midway through development, a fifth researcher joined to implement the mobile client. Implementation fell behind schedule and the researchers made value tradeoffs concerning features to cut. Various core and auxiliary features were cut in the last weeks of development, including automatic transactional saving and integration with existing office applications. Cuts were made carefully avoiding features that we believed would compromise our ability to test the main hypotheses of the project. The final prototypes are shown in Figure 3 and Figure 4.

Study Design and Execution

Concurrently with the research scoping meetings, the research team deliberated on an evaluation approach for Jourknow. The two study types we considered were laboratory and longitudinal evaluation. A laboratory study would have allowed us to directly examine particular aspects of the interface, such as the design of the Pidgin language or the facet panel, whereas the latter (what Kelley and Teevan term a combination of *longitudinal* and *naturalistic* studies [20]) would give us feedback on the integration of the tool with users' lives.

We viewed a laboratory study as too artificial and controlled to be able to reveal how Jourknow might be used to capture "real" information scraps in "real" situations. Furthermore, our previous reviewer feedback indicated a need for longitudinal evaluation of the system. We thus opted for a longitudinal study to give Jourknow a chance to integrate itself into our participants' information management practices so that later we could observe its impact. Our decision carried an implicit assumption that Jourknow would achieve basic uptake, and thus that real-world observation of its research features was a useful next step.

We recruited 14 participants from our university (ages 18-41, median:26), external to our research group. 7 were students at the business school, 1 was visiting Computer Science faculty at our university, 2 were undergraduates and 3 were graduate students in computer science. There were 10 men and 4 women. We randomly divided the group into seven participants who received just the desktop version of Jourknow, and seven participants who received both the desktop and the mobile version of Jourknow (MiniJour). We chose this division in order to perform a between groups study investigating the inclusion of the mobile client, and thus to investigate its effect on take-up of the tool. Both groups had the context capture and Pidgin elements of Jourknow enabled.



15:20 Semantic web: Privacy And per...
09/12 Kristopher's num 6172225652
09/12 @todo meet mark for dinner
09/11 Future directions. Lifeblog
09/11 Interactionlog
09/11 Zipcar reimbursement
09/11 Hyper-Metropolis!

Figure 4. The JourMini mobile phone client.

Following standard practice (*e.g.* [16, 26]), we installed Jourknow on participants' computers, and instructed them in the use of the interface. We also described several of the shortcomings of the current version of the research prototype -- slow loading and saving, occasional GUI bugs, and a remaining server bug that was patched near the beginning of the study. Participants were instructed to introduce Jourknow into their everyday note-taking practices, and to make extra effort to use the software to capture their thoughts and notes. They then used the Jourknow client for a period averaging eight days, including one weekend. Throughout the study, we used e-mail announcements to promote use of the tool, remind participants to integrate the tool into their lives, and keep in constant contact. This level of contact was fell short of other studies which made regular visits to participants (e.g., [8]), but was more direct than those with no reported communication during the study (e.g., [29]).

STUDY RESULTS

Mid-study warning signs. We began to receive indications midway through the study that participants were not making regular use of Jourknow. On the 6th day of the roll out, we observed that only four of the seven participants with mobile phones had tried synchronizing their notes on the server. In response to an e-mail suggesting everyone synchronize, two participants e-mailed us admitting that they had not yet opened the tool, with a third participant experiencing trouble starting the tool on his computer. We helped the participant debug the problem, then sent an e-mail reminding all participants that we had asked them (as per the study agreement) to make daily use of the tool.

Usage analysis. At the conclusion of the study, three participants (P2, P3, P11) had never launched the client after their initial installs. A fourth participant only used the client once right before his exit interview (P1). Others' usage varied significantly. As can be seen in Figure 5, most participants created notes on the day that they received the client, and note creation tailed off sharply in time. Usage picked up when we released a major software patch, and asked users to re-start their clients. Another short jump occurred on the 12th, most likely in response to an e-mail that we sent to participants reminding them that the study was half over, and to remind them to "continue using the client."

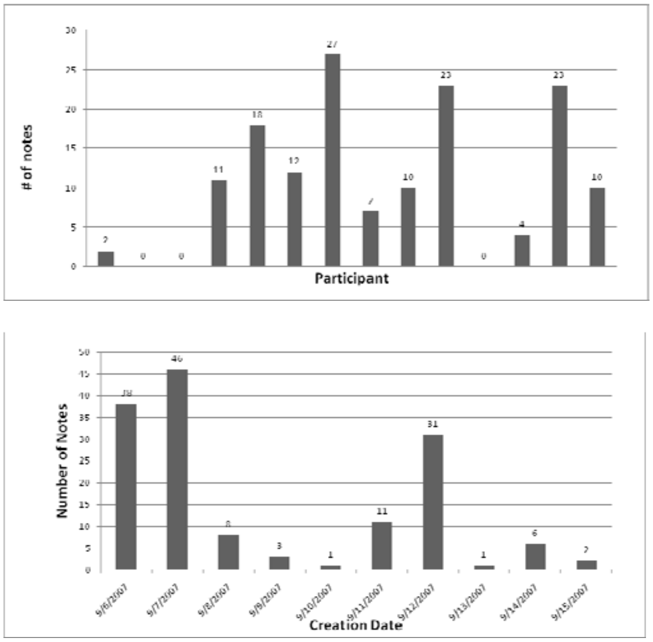


Figure 5. (Top) number of notes recorded per participant. (Bottom) number of notes recorded per day.

Closing interview feedback. During the closing interview, we scheduled each participant to spend an hour with two investigators (one acting as facilitator and one taking notes), where we planned to have participants first provide their general impressions of using the system, and then to walk-through the notes they took using the system, to allow participants to recount their experiences with it.

It took little time to discover that this protocol would need to change due to anemic tool adoption. With our first exit interview (participant 1), we discovered that he had not used the system at all during the week, and had only 2 notes (one of which was created on the day of the install, and one created on the day of, and shortly prior to his exit interview). When asked why he had not used the system despite requests and the terms he had agreed to in the study, he responded “It didn’t become part of my routine, I had to be conscious of it; I’m not accustomed to doing this kind of thing, and it required too much effort for me to bother with it.” Other participants who did not use the tool responded similarly; adopting the tool seemed to require more effort than they wanted to invest. Participant 9 had a slightly different explanation of why he didn’t adopt the tool: “Your tool is just not useful to me. You said that this tool was designed to help people whose ideas just ‘pop’ into their heads, who need a place to write them down. Well, it occurred to me that this just never happens to me! Either I have a lot of ideas that are just not worth writing down, or I just have one good one that I hang on to [in my head] and I don’t need to.”

A majority of the remaining feedback we received focused on highly specific, particular characteristics of the system and of the user interface that they did not like, found annoying or “broken”. These included synchronization “just not working”, complaints about lengthy save/load/launch times, various note views “not working” and being confusing, frustration from the rendering, and issues with ordering and presentation of notes, including font and icon sizes.

Feedback was also occasionally positive, but often inconsistent. Several participants reported liking features (such as the ability to keep notes on the desktop) but was not clear that they had actually ever used it (as they were unclear about how it worked); we also noticed that two participants contradicted themselves by first saying they liked something, and then saying they were annoyed by it or "couldn't stand it" in another context. We received positive feedback about the mobile client, especially those who had phones with QWERTY keyboards; several participants (e.g., p7, p5) reported strongly liking the mobile client running on their phones devices and synchronization logs showed that their devices were actively used. Those with typical 12-key keypads found the mobile client much less useful.

After all of the negative and inconsistent feedback regarding the desktop client, we were surprised when 3 participants protested when began to delete the system from their computers. This was the strongest evidence we had that some participants had actually started to adopt Jourknow into their organizational practices.

Limitations in Our Results

Our user study results exhibited a small number of generalizable characteristics:

Participants' inability to articulate their critique. Whereas we intended to probe for feedback on the general design of our tool and on our research hypotheses, our users were unwilling to provide much feedback on them. Instead, we received very general responses, characterized by broad generalizations such as "I didn't *get* it" or "I didn't find this tool useful." When pressed for reasons, participants (unable to articulate the causes of their disposition) usually paused briefly and then produced a reason which we believe constituted the first plausible justification they thought of. The range and types of reasons varied largely (as described earlier) but largely surrounded overly specific details, failing to provide any larger insight regarding the tool's design.

Inconsistent feedback. When appraising the usefulness of various features of the tool, we often found both inter-participant and intra-participant disagreement. While the former could be explained by differences in individual preferences and practice; the latter, self-contradictions, are troubling -- it suggests that people's appraisals were less reliable as a source of information regarding whether they would truly use the features being appraised.

Lack of adoption of the tool. We observed very little use of our tool amongst our participants. We had requested that our participants insert the tool into their everyday practice, but it was clear that existing practice proceeded with little effect by our tool. Several participants barely used Jourknow during the study period, and several more tried briefly and then ceased to use it.

Lack of coverage over users' varying habits. Though we dedicated a large amount of engineering and design work to covering the basic needs of the information scrap space, we were nonetheless unable to satisfy many of our users. In addition, we received seemingly inconsistent feedback that basic features, while critical to some participants' happiness with our tool, were highly undesirable to others.

REFLECTION ON PRACTICE – WHERE DID WE GO WRONG?

From the above discussion, we see that despite strong momentum going into the final study, the study was unable to test our desired hypotheses. Our goal in this section is to examine the various choices made in the process of designing and developing the first prototype of our system for study. We first discuss why the "obvious" solutions the user-centered design process suggests may not have helped. We then propose four candidate moments for review, what we call *breakpoints* in the process. The **first** of these breakpoints, starting from the beginning, is the original research scope. The **second** is the interface design and prototyping process. The **third** is the type of study chosen, given the state of the prototype, and the **fourth** is the selection of the study population, and the rationale for that choice. For each of these particular breakpoints, we also want to reflect upon how the particular domain of study - personal information management - played a role in our approach. Our process here is to reflect on the methodology that informed our actions in each of these phases, and to investigate what other practices might have better informed our approach.

Considering the "Obvious" Solutions

The most straightforward critique of our process may be the process itself -- that we did not adequately follow the user-centered design mantra. In this section we discuss what might be seen as the most obvious or immediate responses to our situation, and why taking that advice retrospectively may not have helped us.

More UI prototyping! One answer to the lack of user adoption might be that we should have carried out more interface prototyping. To be sure, such lo-fi and hi-fi prototyping would have revealed errors and missteps, for example to improve the visual representation and layout of the facet panel, and structure of the pidgin syntax. However, this prototyping may not have addressed the fundamental issues our study participants reported. For example, our business school participants almost unilaterally did not want to use a computer to take these kinds of notes; if they did, they needed it to be an extension of Outlook, not a separate tool. As we discuss in the breakdowns to follow, our prototypes may have simply been focused on the wrong aspects of the experience.

More system testing! Much feedback we received surrounded participants' perception of the client being buggy and too slow/unresponsive. We have no doubts that more time would have allowed for greater integration and performance testing using more client workstation configurations; which would have uncovered problems that could have lessened this perception. However, it is not clear that even testing our system until it was "perfectly robust" would have received substantially greater adoption, based upon feedback from the couple users who persevered through the glitches and still found many aspects of the tool useless. This suggested that the most important troubles with Jourknow were design-oriented, and that perhaps the glitches were partially a proxy to blame for these more latent underlying design problems.

More iterations! Assuming we had more time, more prototypes, and multiple rounds of "quick and dirty" feedback, the next question is: would our methodology have supported us then? In deference to the "wicked" nature of this problem, the answer is not clear. Why did we see fit to move from hi-fidelity prototypes to a first client implementation? Design is a process of exploration and then refinement [13]; we had refined a prototype that was somehow locally optimal (based on positive informal feedback) but not globally so. Specifically, having employed multiple methods, from interactive sketches, lo-fidelity prototypes to hi-fidelity prototypes, our team felt that we had enough design feedback to proceed with an implementation. Our study results uncovered this error.

Breakpoint 1: Scope of the Investigation

We planned our research to introduce a single tool to address the problems of information scrap capture and retrieval. Our approach to building this tool specified four pillars of design to meet the challenges we had identified in our research: a general note capture and manipulation interface, context capture to facilitate note retrieval, a lightweight structured data capture language (Pidgin), and mobile capture and access. In hindsight, we might ask: did we really have one idea, or four? Should each of these have been studied individually, or were they simply too codependent to do so? What gave us confidence that we could design, develop and evaluate them all together?

At the time of this breakpoint, there were two main factors that played into our decision: the power of the Gestalt in PIM, and positive feedback and inertia from our previous prototypes. We analyze each in turn.

Personal information management tools are such multifunctional devices that they necessarily encompass an entire ecology of use rather than a single research or design problem. This situation leads to two results: huge functionality requirements (resulting in large start-up design and implementation costs), and perception of the system as a Gestalt rather than as singularly differentiable features. Bellotti et al. describe one PIM application, e-mail, as "a mission critical application with much legacy data and structure involved in it" -- and go on to report that several of their users dropped out from the study due to limitations of their research system to adapt to users' complex usage habits [8]. Kelly and Teevan conclude that PIM prototypes must be more robust than typical research prototypes [20], and with both TaskMaster and Jourknow we also see that these tools must also support broad functional requirements in order to compete.

This situation placed us in a difficult position: the system as a whole may not be useful unless we solved several problems simultaneously. Specifically, our inclusion of the mobile client was a response to strong motivation in our previous studies suggesting digital tools severely limit their own usefulness by being available only on a user's workstation or laptop computer. However, in retrospect, the inclusion of the mobile client may have contributed to a prototype unable to anticipate the broad functional requirements supporting our ideas. It is thus questionable whether broadening our scope improved the situation, or simply left us unable to do any of the ideas justice.

A second factor in our decision to incorporate all four ideas into our design was the very positive response we had received from outside reviewers inspecting our work and ourselves. We implicitly took such feedback as design approval and cut down on usability studies of the client. We mistook expert inspection feedback for user feedback. In the space of personal information management, we also see that inspectors may have had difficulty projecting themselves into the use of the client, leading to overly positive feedback.

Breakpoint 2: Prototyping and Interaction Design Process

In designing a complicated system like Jourknow we faced a number of interaction design challenges. Here we examine some of the potential design missteps we may have made, including too few iterations and difficulty prototyping the experience rather than the interface.

The negative feedback we received on the basic design of some pieces of our interaction points to a need for more formative evaluations, earlier on in the process. Design reviews and adherence to precedent were insufficient in our case. One possible solution may have been to use formative laboratory studies during implementation to investigate features in isolation before the longitudinal summative evaluation, or to have lab partners use half-functioning versions of the prototype for feedback.

Our prototypes also faced a challenge simulating the true "experience" of recording an information scrap, rather than simply the interface design. This means that our prototypes succeeded at getting feedback on many interface design challenges, but were less successful at placing that interaction in a context of use. This effect may also have been amplified by our position in the personal information management space, where even small details can make impressive differences in behavior [30]. Our prototypes focused on the novel features -- on being able to re-find information based on context and capture structured information with little effort. Here, we question whether our prototypes were truly effective *experience prototypes* [12], garnering feedback on the rich context surrounding notes' capture and context surrounding reuse. If we failed to prototype important parts of the experience, it is not surprising that user feedback concentrated on unexpected areas of the system.

Breakpoint 3: Study Methodology

The choice of population implicitly assumes the question of the choice of study form: UCD promotes the use of multiple methodologies for evaluation, and recognizes the tradeoffs of different methods in evaluating an interactive system. A point of reflection: was a longitudinal use study the best choice for Jourknow at its current stage of development, and could we have organized the study more in support of our goals?

To recall, we chose a longitudinal evaluation to give Jourknow a chance to ingratiate itself into our participants' practice, and to reflect on how that practice, once engaged, was or wasn't successful. We may now step back and ask: was this decision optimal? Should we have adapted or combined longitudinal and first-use study methods, rather than using them in their typical formulation? For example, we might have begun with a shorter longitudinal study (2-3 days) to identify pain points with the application and then proceeded to use laboratory evaluation to further understand the results. It was potentially to our detriment that we chose the most ambitious study to begin with.

Given that we chose a longitudinal evaluation, did we design the study in such a way as to maximize our chances of success? For example, we chose to keep in contact with participants via e-mail rather than requiring further in-person interviews during the study. Plaisant and Shneiderman [23] and Bellotti *et al.* [8] report that their longitudinal efforts benefited from reappearances to remind each participant of processes in the software that he or she had forgotten about. In previous longitudinal studies of software, however, we see that researchers do often follow up remotely [16, 27] with success. In our case, keeping in close contact with participants would have increased social pressure to use the tool and allowed us to provide follow-up training; this is evidenced by our mid-week e-mail reminder causing a temporary spike in usage.

Breakpoint 4: Choice of Study Population

The quest for external validity [22] dictates that researchers and practitioners randomly choose participants from the target population, rather than form a hand-picked subset. Recently this issue was brought to a head by Barkhuus [6] with a call-to-arms for SIGCHI to stop using local participants (particularly HCI graduate students) in their studies. Thus, pressure from the CHI community to use a random population was a large motivator in our decision to give Jourknow to a group of consisting largely of business students. Here we examine our choice to follow this desire to achieve this new CHI goal for studies to get out of one's back yard rather than testing on participants closer to the research project, or even ourselves.

In the domain of personal information management, ironically, there are reasons why testing outside a friendly community might hurt a study. Kelley and Teevan point out that recruiting PIM system evaluators is a particularly thorny issue: participants must be willing to grant access to personal information, overcome self-consciousness of messy practice, agree to a large time commitment, and commit to temporarily suspending their deeply-ingrained practices [20]. Kelley and Teevan also note that studies in this space, including Bellotti *et al*. [8] and our own, suffer from a degree of participant mortality (drop out prior to the conclusion). A third possible problem lies in community practices in PIM (for example, business students using Outlook) previously unknown to the experimenter. Finally, again due to the "mission critical" aspects of PIM, there is little room for error -- while business students were excellent critics of the system, they were also unable or unwilling to overlook entry barriers to using the system such as outstanding bugs and performance issues.

PIM researchers are left with three main options, then: continue to pursue externally valid studies with outside participants, use insider participants who may be more pliable and willing to evaluate a system through its defects, or "eat your own dog food" and have the researchers themselves reflect on using the system themselves for a period of time. Jones [18] promotes this final option of self-study as a particularly useful tool in PIM research. However, the closer the study population to the research team, the less external validity the results carry. In our case we believed our tool was ready to demonstrate an improvement to a general audience, but this may have been a heavy investment with little return.

POSSIBLE OUTCOMES FOR THE USER-CENTERED DESIGN METHODOLOGY

While we have considered above how we might address next steps for our own process in this project, our overall goal here has been to reflect upon the methodological path or choices that lead us to the decisions we made and produced the results we achieved. Based on this experience, we suggest that the level of certainty various design methods instill in the practitioner or researcher may vary depending on the problem domain. Particularly in "wicked" domains [25] such as information scraps and personal information management, applying a plurality of methods gave us false security that were prepared to build and evaluate a full research prototype -- when in fact basic design elements of our system were still faulty.

We would suggest -- though this proposal itself will need to be validated in some way -- that the breakpoints we have identified in our process may indeed be generalizable breakpoints for others (a) working in PIM research in particular, (b) focused on "wicked" design problems, or simply (c) using multiple methods in any artifact design. We must interrogate the process, and watch for warning signs that indicate a false positive. In our case, our experience prototypes did not succeed in eliciting feedback on the full range of the experience of using our tool.

CONCLUSION

This case study examines the process leading up to unexpected results from a longitudinal usage study of a tool we designed and developed, questioning how a well-grounded approach to test hypotheses derived results that revealed many findings other than those the study was designed to test. Informing parts of this case have been the (a) development of the hypotheses and consequent artifact to be tested, (b) the rationales for the selection of the study participant population, (c) the related choice of study methodology, and (d) the prototyping process for the interaction design. These effects may have been amplified by our chosen domain of information scraps and personal information management.

We have endeavored to show that there is no patently obvious single reason why the study delivered such unexpected results; we have also seen that there is no one clear cut approach to take now that will let us test the hypotheses we wish to test. We propose therefore that this inability to use existing methods to define the "right" path, combined with the numerous unexpected findings from the study, indicate that we need to rethink the design method itself and engage in design of the design, or meta-design. In effect, our use of heuristics (*e.g.*, "If this type of problem, then Method A or Method B work well; choose Method B if time is short, choose Method A if more participants are available"), or reliance on existing methods to cover our design space effectively, stopped us, ironically, from engaging in this meta-design process.

Right now, the most straightforward heuristic would indicate that we have a design challenge outside the scope of standard practice: a new kind of "wicked problem" [25]. Had we focused on our population, study design, scope, and prototyping process earlier on, then we may have taken notice of indications that we had a wicked research problem on our hands rather than something tractable by traditional means. It may be that only surprising results like ours would lead researchers to an investigation of the process. Perhaps we need to reflect on why there are so few of these wicked research problems as precedents in our literature. In the meantime, we shall continue investigating the innovative methodological approaches necessary to accommodate the study of such beasts.

REFERENCES

1. 3M Digital Post It Notes. http://www.3m.com/
2. Circus Ponies Notebook. http://www.circusponies.com/
3. Google notebook. http://www.google.com/notebook/
4. Microsoft OneNote. http://office.microsoft.com/en-us/onenote/default.aspx
5. Stikkit. http://stikkit.com/]
6. Barkhuus, L. and Rhode, J.A., From Mice to Men - 24 years of Evaluation in CHI. in *Extended Proceedings of CHI 2007*, (2007).
7. Bellotti, V., Dalal, B., Good, N., Flynn, P., Bobrow, D.G. and Ducheneaut, N., What a to-do: studies of task management towards the design of a personal task list manager. in *Proc. CHI 2004*, (Vienna, Austria, 2004), ACM Press, 735-742.
8. Bellotti, V., Ducheneaut, N., Howard, M. and Smith, I., Taking email to task: the design and evaluation of a task management centered email tool. in *Proc. CHI 2003*, (Ft. Lauderdale, Florida, USA, 2003), ACM Press, 345-352.
9. Bernstein, M., Van Kleek, M., Karger, D. and schraefel, m. Information Scraps: Eluding Our Personal Information Tools. *In Submission to Transactions on Information Systems*, *http://eprints.ecs.soton.ac.uk/14231/*.
10. Bernstein, M.S., Van Kleek, M., schraefel, m.c. and Karger, D.R., Management of personal information scraps. in *Extended Abstracts CHI 2007*, (New York, NY, USA), ACM Press, 2285-2290.
11. Blandford, A.E. and Green, T.R.G. Group and Individual Time Management Tools: What You Get is Not What You Need. *Personal Ubiquitous Comput.*, *5* (4). 213-230.
12. Buchenau, M. and Suri, J.F. Experience prototyping. *DIS 2000*. 424-433.
13. Buxton, B. *Sketching User Experiences*. Morgan Kaufmann, 2007.
14. Campbell, C. and Maglio, P., Supporting notable information in office work. in *Proc. CHI 2003*, (New York, NY, USA), ACM Press, 902-903.
15. Czerwinski, M. and Horvitz, E. An Investigation of Memory for Daily Computing Events, 2002. citeseer.ist.psu.edu/czerwinski02investigation.html
16. Dumais, S., Cutrell, E., Cadiz, J.J., Jancke, G., Sarin, R. and Robbins, D.C. Stuff I've seen: a system for personal information retrieval and re-use *Proc. SIGIR 2003*, ACM Press, Toronto, Canada, 2003.
17. Hayes, G., Pierce, J.S. and Abowd, G.D., Practices for capturing short important thoughts. in *Proc. CHI 2003*, (New York, NY, USA), ACM Press, 904-905.
18. Jones, W. and Cronin, B. Personal Information Management. in Anonymous ed. *Annual Review of Information Science and Technology (ARIST)*, Information Today, Medford, NJ, 2007, 453-504.
19. Kalnikaite, V. and Whittaker, S., Software or wetware?: discovering when and why people use digital prosthetic memory. in *Proc. CHI 2007*, (New York, NY, USA), ACM Press, 71-80.
20. Kelley, D. and Teevan, J. Understanding What Works: Evaluating PIM Tools. in Jones, W. and Teevan, J. eds. *Personal Information Management*, 2007.
21. Lin, M., Lutters, W.G. and Kim, T.S., Understanding the micronote lifecycle: improving mobile support for informal note taking. in *Proc. CHI 2004*, ACM Press, 687-694.
22. McGrath, J.E. Methodology matters: doing research in the behavioral and social sciences. *Human-computer interaction: toward the year 2000*. 152-169.
23. Plaisant, C., The challenge of information visualization evaluation. in *Proc. AVI 04*, (New York, NY, USA), ACM Press, 109-116.
24. Rettig, M. Prototyping for tiny fingers. *Communications of the ACM*, *37* (4). 21-27.
25. Rittel, H.W.J. and Webber, M.M. Dilemmas in a general theory of planning. *Policy Sciences*, *4* (2). 155-169.
26. Rodden, K. and Wood, K.R., How do people manage their digital photographs? in *Proc. CHI 2003*, (New York, NY, USA), ACM Press, 409-416.
27. schraefel, m.c., Zhu, Y., Modjeska, D., Wigdor, D. and Zhao, S., Hunter gatherer: interaction support for the creation and management of within-web-page collections. in *Proc. WWW 02*, (New York, NY, USA), ACM Press, 172-181.
28. Sellen, A., Fogg, A., Aitken, M., Hodges, S., Rother, C. and Wood, K., Do Life-Logging Technologies Support Memory for the Past? An Experimental Study Using SenseCam. in *CHI*, (2007).
29. Smith, G., Baudisch, P., Robertson, G., Czerwinski, M. and Meyers, B. GroupBar: The TaskBar Evolved *OZCHI*, 2003. citeseer.ist.psu.edu/article/smith03groupbar.html
30. Van Kleek, M., Bernstein, M.S., schraefel, m. and Karger, D.R., GUI--Phooey! The Case for Text Input. in *To appear in Proc. UIST 2007*.
31. Yee, K.-P., Swearingen, K., Li, K. and Hearst, M., Faceted metadata for image search and browsing. in *Proc. CHI 2003*, (New York, NY, USA), ACM Press, 401-408.