Patrick Herron
INLS 150 Organization of Information
Tool review
Submitted 03 March 2005

# *We Can Remember It For You Wholesale*: An evaluation of Google Desktop as a Personal Information Management tool


"Google Desktop Search is how our brains would work if we had photographic memories."

(from http://desktop.google.com/about.html)

## I. Evaluation Criteria

The phrase "Personal information management" (PIM) seems to have two overlapping but distinct senses in the way it is used in the domain of knowledge management. One sense seems to emphasize "personal information" while the other appears to place emphasis on "personal." While the former sense suggests management of information of a private nature, the latter underscores the idea of personalization applied to any information management scheme without a central theme of privacy. In any assessment of a personal information management we must embrace rather than resolve this ambiguity and expect any PIM tool to enhance the use of information that is particular to an individual and at the same time protect the privacy of that person's information. More simply stated, a PIM tool should aid in the organization, management, and use of personal information—information that I use—while helping to protect privacy.

> *"Personal KM is focused on helping an individual be more effective -- to work better."* *(from http://en.wikipedia.org/wiki/Personal_knowledge_management)*

Improving a person's effectiveness on some level seems a reasonable enough requirement for a personal information management tool, but in an important sense the word "effective" is not the exact word when it comes to me. If I am to analyze the efficacy of a PIM tool, I am necessarily limited to such a tool's efficacy *for me*, and *for me*, efficacy is *not* quite efficacy. My requirement for any PIM tool is much simpler and more direct than making me more effective; my goal for a PIM tool is that it should help reduce the amount of time I spend locating information, whether it is information that I have seen before and cannot recall where it now resides, or whether it is information about which I have no recollection of ever encountering whatsoever. For items I know I've encountered in the past, any desktop search tool that fails to have 100% recall will merely slow me down by forcing me to manually look for items. If I know I have it somewhere, I'll persist in searching for it manually until I have found it, despite how grouchy the length of time required may render me. Precision, however, can be sacrificed to some degree. I don't mind sifting through search results if I am confident what I want is somewhere in the list of results, and that list of results is somehow tractable.

One goal in using any PIM tool is to help me search through and find any particular data from the huge pool of data I've accumulated. Not just Word documents or Excel files but also any web pages I've downloaded in the past, emails I've sent, shared, or read. In other words, it needs to be able to help me find documents given my information discontinuities. And it needs to integrate well with the ways I already search for information.

Essentially when it comes to files on my personal computer I do not delete much. I do not feel a need to delete much given the incredibly low price of hard disk

storage and the relatively higher cost of manually finding disposable data and disposing it. That's not to say I don't delete emails I will never have any need to read again to wiping out applications and related data I no longer use—I'm aggressive about ridding my machines of what I'm certain never to use again. While I do not save everything, I do err on the side of saving, choosing to save everything I will have any remote chance of using again. At home between my main workstation, a backup machine and workstation for my wife, a Linux box that I "mess around" with, and an old laptop, I have about 600GB of storage, and maybe 480GB of that space is taken up. When adjusting for redundancies due to RAID and backups, I perhaps have about 300GB of data, and perhaps about 200GB is data that I've created or am responsible for in some way: from cached web pages to scholastic homework assignments to programming code I've written to my email archive (approximately 7000 emails) to my collection of writings (approximately 1500 original poems with an average of 3-5 revisions per poem) to backups of web sites I have authored to original graphics files for print and web design projects to my family's digital photograph collection to mp3s to tax records, business correspondence, legal affairs, and so on. In short, I have a lot of personal data, and it is highly heterogeneous, both in terms of file type and in terms of concept, purpose, and task.

> *"It assists in discovering, aggregating, relating, and management of the information and knowledge needed to do his/her own job"* (from http://www.voght.com/cgi-bin/pywiki?PersonalKm)

Well, what is my job exactly? Or rather, what are my jobs? My regular tasks? Well, foremost, I am a full-time student at UNC-Chapel Hill in Information Science. My studies require work across heterogeneous computing environments. My school work and class-related computing work is performed on two remote *NIX servers (jade.ils.unc.edu and baobab.unc.edu) as well as on my IBM laptop (Win) as well as my home workstation (Win). I maintain an art website on a commercial server, and that domain space also contains some other

web-published data, such as cached news articles, links to previously published content, a professional resume, a CV for my creative work, and even some school class presentations. I also work as a publishing poet, an activity that has two main components: writing poetry and sending it out. The writing task for me is hardly one that necessarily operates on ideas materializing as if out of thin air; my approach to writing poetry over the last eight years often has at times and in particular projects been heavily dependent upon manual NLP-type tasks performed on extant texts followed by often-aggressive editorial policies: a process of procedural traduction and refinement. I depend heavily on the "background noise" of the internet for the starting material: anything from famous poems to absolutely terrible poems to newsgroup discussion and even spam. Sometimes I even use personal information already on my machine. In this case, high precision searches are in some way counter-productive for me, as I want some element of surprise in my searches, surprise that I imagine would be all but eliminated from high-precision results. I also run a business, currently a small contract with a single client; I work as a research assistant at ibiblio; and I also have a local public function that includes running a public literary festival. This work is spread across servers and personal machines.

Importantly, there's no strict division of data on these remote environments. Sometimes my text mining research data ends up on my ibiblio workstation or on my laptop. Sometimes my school presentations are posted in the domain space where my creative work resides. My writing is collocated on my personal domain server as well as on my personal work station and my laptop. In other words, my personal information cannot be easily classified with respect to task by the machine on which it resides.

Physically speaking I do all of my work at three different machines; other machines that I work on, I connect to from these three machines. The three machines are my home windows personal work station, my windows IBM

workstation, and my ibiblio Linux workstation. Work on all other machines is performed using a ssh client. It would be very nice if a PIM tool could record my session histories.

What O wpould really love is a PIM tool that could separate or cluster my personal information according to specific types of tasks I perform:

- School
- Text mining research
- ibiblio
- poetry / writing
- Art / design
- Managing poetry festival
- Business
- General research (literary & political)

And these tasks might be generalized as follows:

- Managing web content
- Academic research esp. literature searches
- Managing CV-type information (keeping track of news stories, publications, projects, etc. for both the academic-professional and the artistic-poetry domains, two domains I strive to keep separate)
- Managing writing and art submissions
- Writing
- Designing

Now, most machines I work on have search functions, from *grep* and *find* on *NIX machines to the built-in search tools on my Windows machines. They don't search my web cache or my emails, for example. And they have no grasp of any image-related data. Further, the search results form these search tools are highly unfriendly and offer only poor browsing at best.

I also have a dominant set of tools I use for searching, and I would find it useful if a new tool could mesh in some way with some of my search behaviors. Like many, I make heavy use of Google to navigate the internet, but I also make heavy

and regular use of ISI's Web of Science, Lexis/Nexis, my OS search tools, and also search engines like Clusty (http://www.clusty.com) and also the search features within iTunes and even my email client. My behaviors with these tools are trivially but importantly personal behaviors that are represented by the information exchanged in these tasks, and that information is personal in that it is relevant to me.

As for organizing my material I am not unlike many of the users described in Barreau & Nardi's "Finding and Reminding[1]": I do not have an elaborate scheme for organizing. And to a large extent, despite my scheme's simplicity, I am failing to keep up with it. On my windows machines I generally have four "buckets": desktop folders into which nearly everything falls. Those folders are labeled "personal", "school", "art", and "professional." Files that haven't found their way to one of the buckets typically either remain on my desktop or inside a folder called temp, also on the desktop. On Linux machines, most files linger in my account login directory, but creating and naming files on *NIX systems for me is a bit more detailed and elaborate than when I'm working on Windows machines.

Finally, with this huge variety of tasks, environments, needs, formats, and established search behaviors, it would be nice if I could bring the work together somehow. Further, the OS's search tool on my Windows workstation has died with no hope for a quick repair in the near future. This puts an even greater onus on a PIM tool, as to some extent it will need to operate as a short-term replacement. It also underscores a desire I perhaps share with many other users: I do not want to spend a lot of time implementing or maintaining this tool.

[1] Barreau, D. & Nardi, B. (1995). Finding and reminding: File organization from the desktop. *SIGCHI Bulletin*, 27(3), 39-43.

I might summarize my evaluation points for any PIM tool as follows:

- Privacy: PIM tool does not make personal data less secure
- Particular to me—what I use
- Adapts to heterogeneous environment – remote *NIX servers, personal Linux boxes, multiple email accounts from different arenas, as well as a large variety of data formats
- Runs on all three machines and records terminal sessions
- High recall but willing to sacrifice a little precision, especially for the sake of the occasional surprise
- Integrates well with current search behaviors (heavy use of OS search tools, Google, ISI Web of Science, Lexis / Nexis, and iTunes)
- Improvement over OS-based search tools (searches contents of files, represents results in easy-to-browse format)
- Good for individual task-based information sets
- Ultra-simple installation & zero maintenance
- One-stop shopping: help unify various search efforts

I realize my criteria are not general but instead specific to me. Despite the seeming narrowness of the scope of evaluation I have constructed, it may be the *only* valid approach to evaluating personal information management tools I have available. Evaluating an automobile yet never actually placing it on a road is unlikely to be a reasonable approach. The road of the PIM is the person-in-context. I have learned from the spirit of the knowledge management literature that the best approach necessarily involves contextualizing as much as possible, and that generalizations may spring forth from the fount of particulars. I am hoping, then, for a little serendipity, in a bottom-up approach.

## II. The Tool: Google Desktop

The people at Google introduce their new Google Desktop Search tool by making a rather far-fetched claim that "Google Desktop Search is how our brains would work if we had photographic memories." After climbing up to this somewhat oxygen-starved height, we subsequently learn from Google that the tool is significantly less grandiose than any claim suggesting its status as a prosthetic hippocampus. Google Desktop, according to Google, is designed to provide full-text searching capability over an individual's email, cached web pages, chats, and other "files." Searching one's own computer is promised to be as easy as searching Google, and you are even supposed to be able to search web pages you have visited before, even if you are not currently online. Importantly, Google promises that the tool will remove any need for actually organizing one's own files, since the representation the tool builds provides the very organization one needs.

Before I explore where Google Desktop does and does not meet my own particular criteria for a Personal Information Management tool, I first want to discover if the tool meets the high-level design goals articulated by Google. Having installed the tool on two of my machines back in December, I'm already fairly certain that Google Desktop has not served as a replacement for my brain, however desperately I feel the need for one from time to time. As for the other claims, I anticipate uncovering a number of positives, heavily-asterisked problems. (I should note here for the record that installation was a breeze—one of the easiest installations ever.)

Google Desktop indeed supports full text searches over email, web history, chats, and other files. Here come the asterisks. Google supports full email searches if you use Outlook or Outlook Express as your email client. I have many email

accounts, but I read only two of those accounts through Outlook, none through Outlook Express, and I am presently in the process of abandoning Outlook altogether in trade for Mozilla Thunderbird, a much more robust and secure email client. Despite the switch, I have over 7000 emails in an Outlook-based file format, and Google's search capabilities over that content is quite powerful. For example, a search on a poet-friend's name, the name of someone with whom I corresponded with extensively, produced approximately 400 emails, and nearly all of which were emails either from me to him or from him to me. Looking in Outlook, I saw there were approximately 350 such emails between us. The other 50 or so results in the Google Desktop search were emails about him or web pages I had viewed in the past with content about him included. It appears that for simple search tasks that Google's email search has superb recall and a sufficient precision that fortunately allows for results I hadn't considered when entering his name as a search term, results that were highly relevant despite my lack of expectation about seeing those results. What's more, the email search results are organized by threads and subjects, and the results can be sorted either by relevance or by date.

Google Desktop currently supports searches only of the following file types:

- Microsoft Outlook & Outlook Express
- Microsoft Word
- AOL Instant Messenger
- Microsoft Excel
- Microsoft Internet Explorer
- Microsoft PowerPoint
- Plain text

Some common file types I depend upon heavily include Adobe PDF files and html files not included in Internet Explorer, such as HTML files I've authored locally, or, perhaps more importantly, HTML files I've browsed online while not using IE (I use Mozilla Firefox exclusively on both of my Windows machines and

my Linux box alike). Google Desktop is not indexing those files for searching. I also use Gaim instead of AOL as an IM client, and Google Desktop is not indexing logs of my Gaim-based IM chats.

Further, not only are the file types limited to that set, but also, Google Desktop at the present moment is solely for use on My Windows machines. No Linux or Mac version is offered at this time.

Does the tool actually *eliminate* my need to organize my files? While it has *reduced* my need for organizing them, it certainly has not *eliminated* it. While creating and maintaining a manual organization scheme for my files is not needed for searching them via Google Desktop, it is needed for browsing; Google Desktop is hardly a robust tool for browsing one's own personal files. For my own browsing tasks, I would be completely lost without my own manually-derived "bucket" scheme I described earlier: my "personal," "professional," "school," and "art" folders. To some extent, however, I must be fair and credit Google Desktop where credit is due. I realize I have learned to browse by searching, and Google Desktop has facilitated that sort of approach, and has as a result reduced the amount of file organizing I perform. When I say "browsing by searching," mean that I use a search tool to browse, and I do that by entering a very general search term. But such broad terms are not powerful enough to reflect the broad task categories I elucidated earlier. For example, a search on "writing" is not going to produce a set of results that approximates the complete set of all materials related to writing. The Google Desktop tool is bound to term dependence.

Searching by Google Desktop is as easy as searching Google. In fact, behaviorally, there's little difference. Google Desktop integrates with Google. That is, once you have installed Google Desktop and have it up and running, whenever you go to Google to do a search, a Google Desktop search will be

performed just as the Google search is being performed. After installation, you will see a desktop option offered (see fig 1) incorporated into the Google search page itself that will lead you to a separate Google Desktop search interface (fig 2) that looks exactly like the Google search interface.

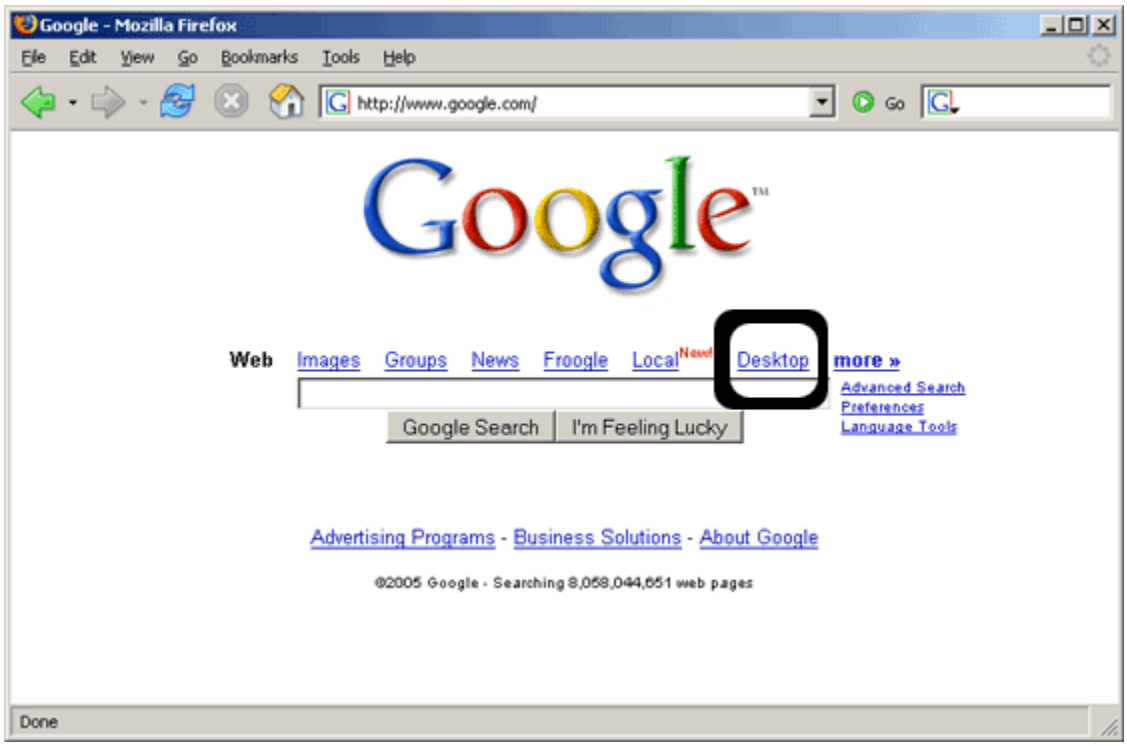


**Figure 1 Google Desktop option integrated into view of Google search page.**

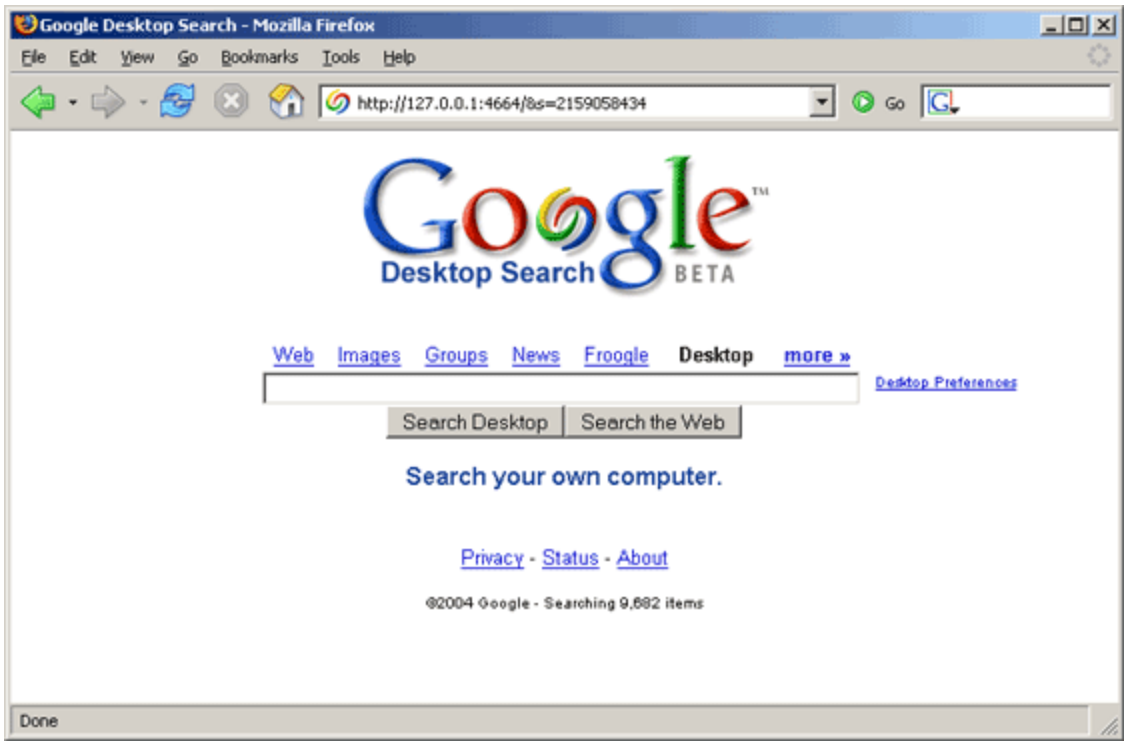


**Figure 2 The Google Desktop search interface looks just like the Google search interface.**

The integration of Google Desktop is better than that, however; you do not even need to go to the separate Google Desktop search interface to use it—you can search your machine directly from the Google search window itself. For example, I enter the term "apple" (fig 3) in Google. You will note that the Google results page is headed by desktop search results first (fig 4), just above the standard web search results.


**Figure 3 – Performing a search in Google after installing Google Desktop**

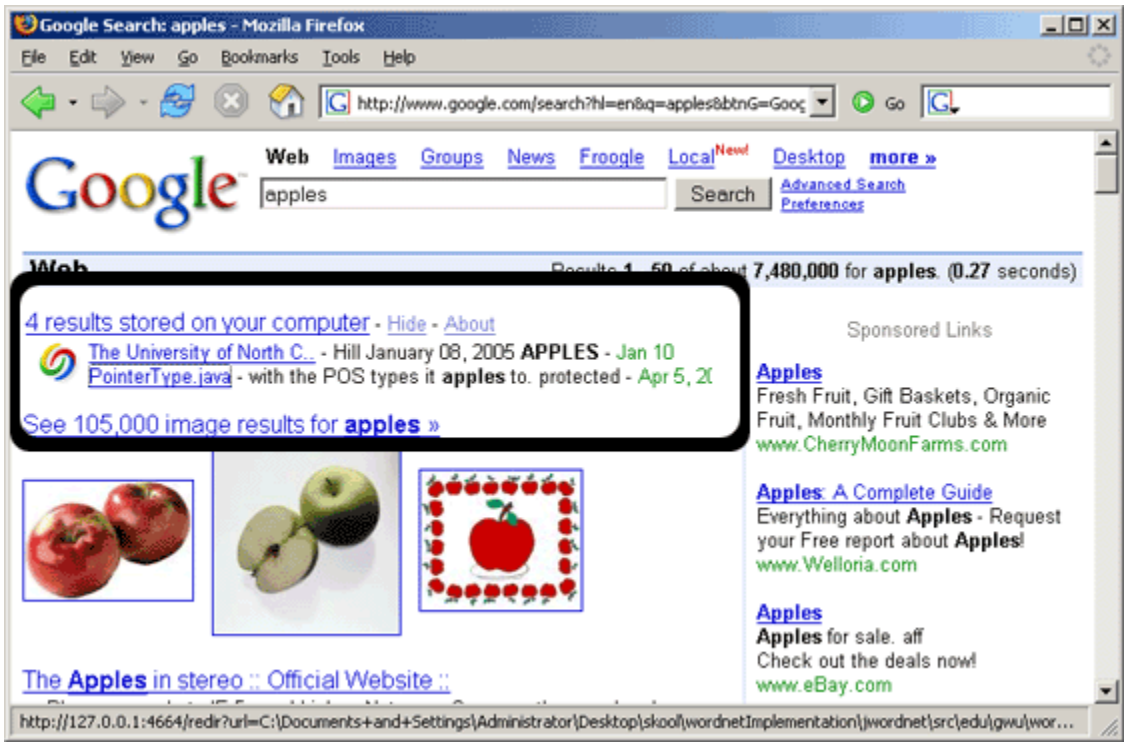


**Figure 4 The standard Google search results are modified to include Google Desktop search results at the top of the standard web search results.**

Selecting “results stored on your computer” leads to a page constructed solely of a list of desktop-only results, in a format that appears exactly like the typical Google search results to which many have grown accustomed (fig 5). However, the Desktop search adds the very convenient feature of providing a thumbnail image of any HTML results.


**Figure 5 Desktop search results look just like regular Google search results, except with the addition of a snapshot thumbnail of HTML results (top right).**

When I started this analysis I began by establishing criteria for evaluation relevant to my own personal needs:

- Privacy: PIM tool does not make personal data less secure
- Particular to me—what I use
- Adapts to heterogeneous environment – remote *NIX servers, personal Linux boxes, multiple email accounts from different arenas, as well as a large variety of data formats
- Runs on all three machines and records terminal sessions
- High recall but willing to sacrifice a little precision, especially for the sake of the occasional surprise

- Integrates well with current search behaviors (heavy use of OS search tools, Google, ISI Web of Science, Lexis / Nexis, and iTunes)
- Improvement over OS-based search tools (searches contents of files, represents results in easy-to-browse format)
- Good for individual task-based information sets
- Ultra-simple installation & zero maintenance
- One-stop shopping

While I have not yet explicitly addressed those items directly, it seems that in analyzing Google Desktop on its own terms—in terms of whether it meets Google's articulated design goals for it—I have fortunately covered a good number of these criteria. The search does record and organize personal data, personal in the sense that it is particular to me. It does run on my Windows machines but not my Linux machine, and its searches are confined to a set of file types that do not represent all of the file types I need, missing things like terminal session data or all web session data, particularly web sessions not run in IE. The searching does have excellent recall and the relative looseness in precision does allow for surprise results. The tool is excellently—perfectly—integrated with one of my most dominant search behaviors. It allows me to browse my personal information through very general search terms and thus pertains well to my individual task-based information sets I have already established. Installation was a breeze. I can even find my mp3s through it, rather than having to resort to opening up iTunes, thus suggesting it does in fact reduce the number of places I need to go in order to find my personal information, and it even reduces the amount of organizing I need to do. I have not yet, however, addressed maintenance or privacy.

One of the major shortcomings of Google Desktop is one I have encountered much to my surprise after using the tool for some time. Google Desktop, it seems, has difficulty maintaining an accurate and up-to-date index of desktop contents. In particular, when I move an item on my desktop machine that has already been indexed by Google Desktop, Google Desktop "loses" it. It appears

that Google Desktop does not update its index to respond to previously indexed files that have been relocated or deleted. When I recently performed a major reorganization of files and folders on my home desktop, I could no longer search for or in any of the files that had been relocated. Anything that is moved is not reindexed to match the new location of the file; the search results will point to the original location of the item when it was first indexed. In some sense, then Google Desktop exerts a pressure on the user to *not* reorganize files, for if the user does reorganize them, the file will be lost from the Google Desktop search. I have found myself doing an uninstall and reinstall of the tool in order to reflect the changes, as there's no "reindex" feature to ameliorate this major flaw. By reinstalling I am forcing the tool to reindex the contents of my hard drive, but this indexing process can take hours and considerable processing power, thus slowing down everything else on the machine. I believe this is a flaw significantly worse than its limiting of web session data to IE.

Since the release of Google Desktop a number of concerns have been raised about its security. Foremost on that list is a concern that a user having a central index repository of his or her web behaviors makes it easy for someone unscrupulous to poach at once a treasure trove of personal information, from emails containing passwords to personal conversations on chats to financial information. For the average windows user who exclusively uses IE, Outlook, AOL IM, and Office it is possible that everything personal on one's computer is centrally available within Google Desktop's data repository. Security, which includes but is not limited to protecting a person's privacy, is after all about slowing down violations rather than eliminating them altogether; eliminating theft is impossible, but the longer the theft takes, the less incentive there is for the thief to take that data.

It appears for some very sensitive data Google Desktop actually makes it easier for the unscrupulous to find it. For example, imagine a scenario where a person

working in an office environment occasionally manages her checking account at work. That person, who just happens to have Google Desktop installed, forgets her password to her bank account website and has fills out a form on the back website to have her password sent to her via email. She then receives and reads the email. One day she walks away from her machine, and someone else opens her browser and searches for the word "password." Voila, high on the search results is an email containing the user name and password to that account. No cracking was necessary in this scenario, no technical wizardry at all—just eight keystrokes and three clicks and a stranger is in her back account routing money out of the account.

Computer criminals could have a field day if able to install Google Desktop on public machines. Google Desktop could make it easy for someone looking to harvest identities or even gather the details of a lurid extramarital affair for the purpose of blackmail. Users who are not aware of the consequences of such a powerful tool place themselves and others around them—notably family and employers—at considerable risk in using this tool. For cybercriminals the tool creates nothing less than a honey pot with a convenient well-integrated interface that's easy for them to use.

Google Desktop is a fascinating and useful tool that lives up to some claims while falling short of some crucial expectations. The easy-to-install too that meshes seamlessly with regular Google searching does provide more powerful searching than typical OS search tools, even recording your web history for you. But the tool is only available for Windows and only tracks in my estimation a small fraction of the data and tools needed to be tracked. Further, the tool doesn't update after a file has been moved to reflect the new location, thus leading to search results with broken links. While it doesn't replace the need for organizing one's personal data, it does reduce some of the pressure of such a need. Unfortunately at this point it is not altogether convincing that even in the

hands of a savvy and cautious user the benefits of the tool outweigh the security and privacy risks. The product is in beta, however, and we should expect some significant enhancements of privacy and security features of the tool before its version 1.0 release.

## III. Revising the Evaluation Criteria

Given the criteria I have derived, I would add to them a set of criteria related to evaluating a tool on its own terms. I.e., does a tool actually do what it promises to do? Simply asking whether a tool meets my needs isn't enough; it is often the case that a person selects a tool to help solve some of his problems only to find that after adopting the tool he or she is addressing and attempting to answer a new set of problems. I find it instructive to consider the design goals as evaluative points because it allows me to be influenced by needs I had not previously realized as my own while also helping me to better articulate how it is a tool is meeting or failing to meet my own needs.