A PE
for

MyLifeBits: RSONAL DATABASE EVERYTHING

Developing a platform for recording, storing, and accessing a personal lifetime archive.

BY JIM GEMMELL, GORDON BELL, AND ROGER LUEDER

The January 2001 *Communications* article [1] "A Personal Digital Store" described our efforts to encode, store, and allow easy access to all of a person's information for personal and professional use [1]. The goals included understanding the effort to digitize a lifetime of legacy content and the elimination of paper as a permanent storage medium. We used Gordon Bell's document archive as well as his current activities as a vehicle for this research. It was presumed that an emerging terabyte disk would hold a lifetime of accumulated information of a moderately active professional person. This article describes the project's progress, insights, and surprises over the last five years. From the original plan of simply storing files of scanned papers, we evolved a concept of what the PC of the future should look like as we developed the SQL-based MyLifeBits platform.

Illustration by Jean-François Podevin

Since 2000, 40GB disks at $10/GB have been replaced by 500GB disks at less than $1/GB, with terabyte drives expected to arrive by 2008. While disk capacity was expanding, so was Bell's digitized life. His non-video content grew at a rate of approximately 0.5 GB/month, but in a nonlinear fashion. Email messages grew larger as attachments became more common, and digital photos took more space each time a new camera with more megapixels was purchased (2 MPixels in 2000; 5 MPixels in 2005). His experience has been consistent with the idea that, absent video, a terabyte still seems adequate for lifetime storage, as a terabyte can hold more than 1GB/month for the duration of an 80-year life assuming only modest storage for what is seen and heard (see the table here for the file formats contained in Bell's store).

Item type	Number	Size (GB)
Video	1,303	61.3
Audio	5,083	12.2
Pictures	43,812	8.8
Tiff	3,832	7.9
Web pages	70,918	5.7
PDF	3,527	4.9
PPT	1,815	4.5
Doc and RTF	13,764	1.2
Other	5,046	0.9
Email text	97,271	0.3
Total	245,371	107.6

* Tiff and PDF hold about 250,000 pages as single and multiple document files

Gordon Bell's content formats circa November 2005.

However, changing user patterns could invalidate this assumption. We now record things speculatively—recording what we might want to see later. Additionally, our effort to "capture everything" moved beyond legacy content like paper, photos, and video, into a second phase that included real-time capture of conversations, meetings, sensor readings, health monitors, and computer activity. In the future, we may start taking 1,000 photos a day (as is now feasible with SenseCams [5]), or storing all meetings and conversations, or storing photos in raw rather compressed form. Inclusion of video can easily exceed 1GB per month or even a day. Indeed, in a brief experiment recording television programs that might be watched we quickly acquired nearly 2TB of material. We now believe a terabyte will hold a lifetime at 20th century resolutions and quantities, but speculate that 21st century users may expect to record their lives more extensively and in higher fidelity—and may drive a market for much greater storage.

The original project avoided the use of a database, using only a file system with careful naming of files and judicious use of folders and shortcuts. However, as the collection grew the use of files in folders went from unwieldy to overwhelming. In 2000, search tools were cumbersome. Current desktop search tools are vastly superior, but they still work in terms of files and folders. We wanted more powerful capabilities, such as access by metadata including written and spoken comments about items, the ability to organize items in multiple ways, and to test different ways to organize and classify information.

Faced with these challenges, the focus shifted from capture to the development of a software platform to make the captured material manageable and useful. The new project that began in late 2001 was dubbed MyLifeBits.[1] We hoped to substantially improve the ability to organize, search, comment, and utilize content. We also wanted to obtain a single database in contrast to the many data "islands" being created including mail, contacts, meetings, finances, health records, photos, and other items. Frustration with the file system led to testing the suitability of databases for personal storage, and ultimately into research about next-generation storage systems.

Memex as a Blueprint

For inspiration, we looked back on Vannevar Bush's 1945 article "As We May Think" [3]. Bush had a strong grasp of American science and technology, having been director of the U.S. Office of Scientific Research and Development throughout World War II, where he "coordinated the activities of American scientists in the application of science to warfare." Two years before the invention of the computer and transistor he asserted that "instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages." His 60-year-old article is a prophetic blueprint that includes the computer, low-cost library storage occupying a small amount of space, commerce with automatic inventory control and billing, fast communication, speech interfaces, and a hypertext-linked cyberspace. Our interest is in his all-inclusive, personal information system, which he called "Memex."

Bush posited the Memex as "a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory." The Memex was to be built into a desk with a keyboard, microphone, and display surfaces. Its interface could copy photos or papers, or could be written on. However, "most of the Memex contents are purchased on microfilm ready for insertion."

[1] The initial project had been called CyberAll, a name that was discovered to already be held by United Services International.

In a typical use scenario, the user "moves about and observes, he photographs and comments. Time is automatically recorded to tie the two records together. If he goes into the field, he may be connected by radio to his recorder. As he ponders over his notes in the evening, he again talks his comments into the record" using speech-to-text. With a walnut-sized, forehead-mounted camera, the user "moves about…every time he looks at something worthy of the record, he trips the shutter and in it goes…".

Bush wanted to improve on the experience of physical libraries, but realized that the problem "goes deeper than a lag in the adoption of mechanisms by libraries, or a lack of development of devices for their use. Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing." An item can only be in one place, and to find it "one has to have rules as to which path will locate it, and the rules are cumbersome. Having found one item, moreover, one has to emerge from the system and re-enter on a new path." Bush pointed out that the "human mind does not work that way. It operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thought, in accordance with some intricate web of trails carried by the cells of the brain." He suggested that items in Memex could likewise be organized in trails. Bush's idea of a "web of trails" is often credited as inspiration for the World Wide Web. However, Memex was a personal device, akin to the PC.

MyLifeBits Software

Memex, with links and comments holding a central role, served as our blueprint for MyLifeBits. Faced with folders full of documents, messages, phone calls, photos, and music files, with inherent or potential metadata such as author, camera, comments, location, and time, we needed a framework to hold and link all of these objects in the web-like and almost arbitrary fashion that Bush described. We deemed search to be the most critical requirement. Furthermore, we realized metadata is often a key part of user recall, for example, that an email message was sent during a certain year, that a song was by a certain artist, or that a photo was taken at a certain place. Holding and linking all these items is exactly what databases do.

MyLifeBits has at its heart a SQL Server database that can store content and metadata for a variety of item types, including contacts, documents, email, events, photos, music, and video (see Figure 1). Currently, our database supports 25 item types. Each item has about 20 common attributes. Additionally, each has a database table. For example, contacts have an additional 62 attributes, including email address and date of birth.

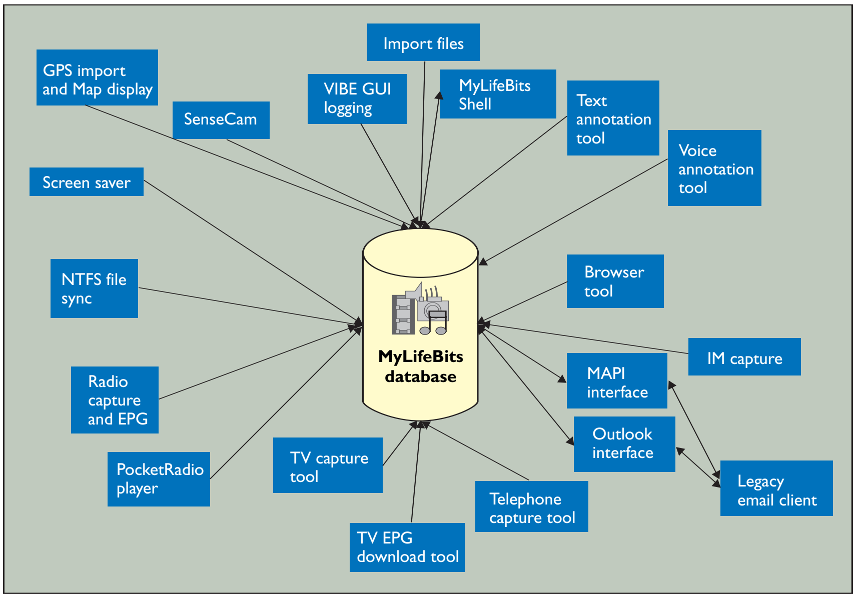


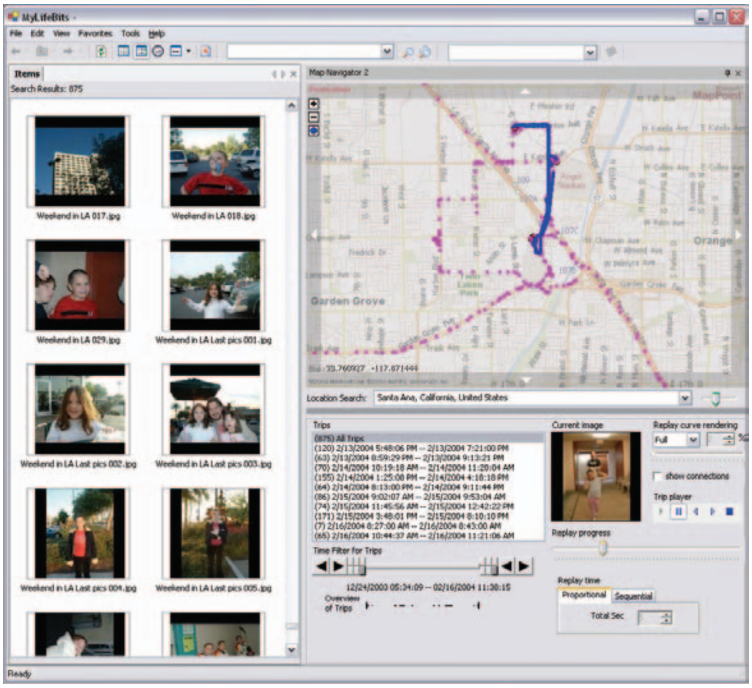
Figure 1. The MyLifeBits platform store and the suite of capture/display tools.

Items can be linked together implicitly using time to "tie" them together as Bush suggested; or explicitly linked with typed links such as a "person in photo" link between a contact and a photo, or a "comment" link between a voice comment and a document. With linking, the traditional folder (directory) tree can be replaced using our more general "collections" function based on using a directed acyclic graph (DAG). Any object (including a collection) may be filed in any number of parent collections.

Metadata and linking are nicely illustrated by their use with photos. We expect future cameras (including cell phones) will automatically label every photo with time and place using embedded location hardware. Thus a photo can be recalled by a label, when or where taken, subject, and so forth. Eventually, we expect software to analyze a photo's content to create additional metadata. At present, our software allows photos to be located by dragging and dropping onto a map to link a location with a photo. Similarly, a "person in photo" link can be used to manually connect a photo and contact. Alternatively, a common timestamp implicitly "ties" the time a photo was taken with a person's GPS recorder location to create location metadata. Figure 2a shows an animated trip log with photos and GPS locations marked on a map. Photos may also be linked to calendar events to indicate a photo of the event, turning the calendar into a photo diary as in Figure 2b.

Once everything is in a database, the project becomes a quest for useful tools to organize, associate metadata, access, and report about the information. Figure 1 shows the functions we have built to date around the MyLifeBits database. In order to support legacy applications, NTFS files and Outlook email stores are monitored and their metadata integrated into the database, including the text of each item to enable full-text search. The system captures every Web page visited, all Instant Message chat sessions, all telephone conversations, as well as meetings, radio, and television program usage as shown in Figure 1. A GUI logger records all mouse and keyboard activity (see Figure 2c). This log can reveal the significance of an item based on use, or can reveal insight into how one spends time with the computer. Office audio/video recording is our most recent capture application.

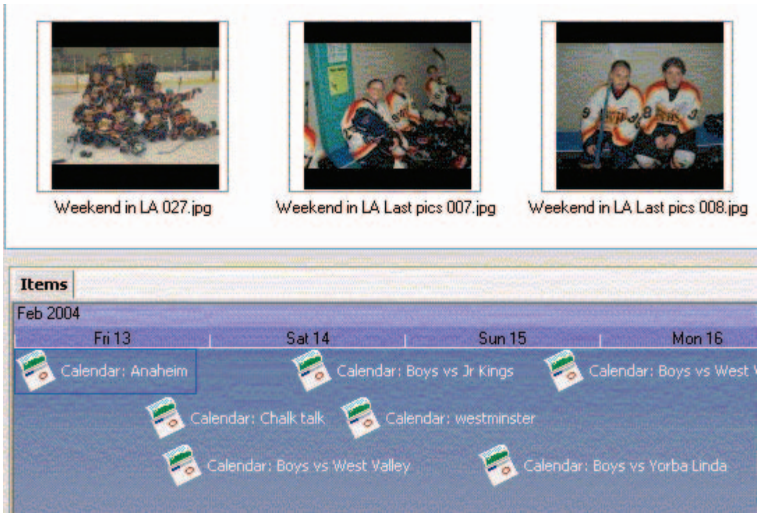
The MyLifeBits shell is the main user interface. It allows queries to be viewed as a list, variable-sized thumbnails, and a timeline. The user interface enables refinement or pivoting according to metadata and links, as we will describe, and provides for the creation of text and voice comments. For example, any number of selected items may be commented on with



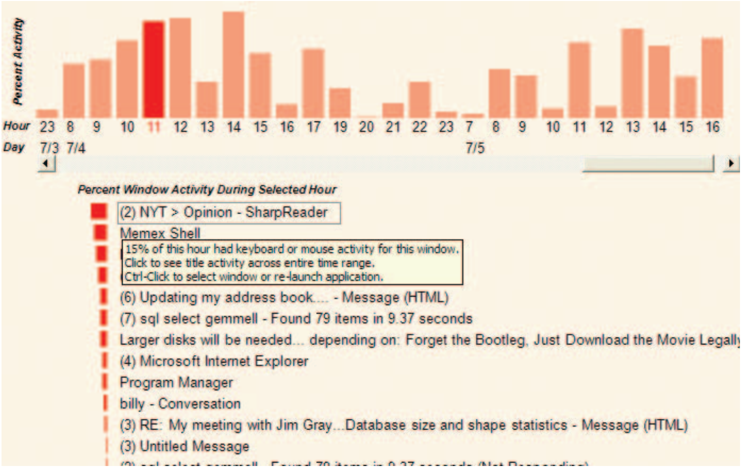
Trips
(875) All Trips
(120) 2/13/2004 5:48:06 PM -- 2/13/2004 7:21:00 PM
(63) 2/13/2004 8:59:29 PM -- 2/13/2004 9:13:21 PM
(70) 2/14/2004 10:19:18 AM -- 2/14/2004 11:20:04 AM
(155) 2/14/2004 1:25:08 PM -- 2/14/2004 4:18:18 PM
(64) 2/14/2004 8:13:00 PM -- 2/14/2004 9:11:44 PM
(86) 2/15/2004 9:02:07 AM -- 2/15/2004 9:53:04 AM
(74) 2/15/2004 11:45:56 AM -- 2/15/2004 12:42:22 PM
(171) 2/15/2004 3:48:01 PM -- 2/15/2004 8:10:10 PM
(7) 2/16/2004 8:27:00 AM -- 2/16/2004 8:43:00 AM
(65) 2/16/2004 10:44:37 AM -- 2/16/2004 11:21:06 AM

Time Filter for Trips
12/24/2003 05:34:09 -- 02/16/2004 11:38:15

a


Weekend in LA 027.jpg
Weekend in LA Last pics 007.jpg
Weekend in LA Last pics 008.jpg
Items
Feb 2004
Fri 13
Sat 14
Sun 15
Mon 16
Calendar: Anaheim
Calendar: Boys vs Jr Kings
Calendar: Boys vs West V
Calendar: Chalk talk
Calendar: westminster
Calendar: Boys vs West Valley
Calendar: Boys vs Yorba Linda

b



Percent Window Activity During Selected Hour

(2) NYT > Opinion - SharpReader
Memex Shell
15% of this hour had keyboard or mouse activity for this window.
Click to see title activity across entire time range.
Ctrl-Click to select window or re-launch application.
(6) Updating my address book.... - Message (HTML)
(7) sql select gemmell - Found 79 items in 9.37 seconds
Larger disks will be needed... depending on: Forget the Bootleg, Just Download the Movie Legally
(4) Microsoft Internet Explorer
Program Manager
billy - Conversation
(3) RE: My meeting with Jim Gray...Database size and shape statistics - Message (HTML)
(3) Untitled Message
(2) sql select gemmell - Found 79 items in 9.37 seconds (Not Responding)

c

Figure 2a. Map interface: pink dots for GPS points, red dots for photos, and a blue line used in animated trip replay.
Figure 2b. A calendar becomes a photo diary when photos are linked to events.
Figure 2c. GUI activity log on an hourly or daily basis (courtesy of George Robertson).

a simple button or right click operation (comments may be text, voice, or any file). Similarly, these items can be assigned to collections. The screensaver displays random photos and video segments, and gives the user an opportunity to comment and rate items. Simple authoring tools create side-by-side timelines and HTML-based slide shows with audio.

Experience and Observations

Having a surrogate memory creates a freeing, uplifting, and secure feeling—similar to having an assistant with a perfect memory. Since we can't predict when an old bill, conference announcement page, attendee list, or business card will be required, the easiest and safest thing is to simply keep it all. Our only store, including financial and legal documents such as bills, contracts, pay stubs, trusts, and wills is electronic. Stock certificates are the only retained paper.

Part of feeling secure is knowing that capture is increasingly automatic. While browser Web capture at first struck us as somewhat trivial, this is an essential feature that has changed our behavior. The failure of one author's hard drive that resulted in losing four months of captured Web pages was a severe emotional blow—perhaps like having one's memories taken away. Even months later, he searches for information expected to be in the Web archive, only to realize it was lost. We routinely visit pages just to ensure having a copy. Undoubtedly, our progeny will wonder why we were there. Our corporate intranet is an important information source that includes forms, documents, and presentations ranging from health insurance to product specifications. As many internal sites are changing and transitory, having a personal copy is essential.

HAVING a surrogate memory CREATES A FREEING, UPLIFTING, AND SECURE FEELING—similar to having an assistant with a perfect memory.

The good news is that more and more content is being "born digital" without the need for scanning. We expect that soon all information will arrive digitally, including bills, correspondence, financial statements, music, and photos. Articles in professional journals, newspapers, and magazines are perhaps the most valuable content that a professional has and these are available digitally now. RSS feeds from professional organizations will improve this situation. This not only implies that the scanning process will be eliminated; there is also the opportunity for metadata to be included at virtually no cost. For example, in the future, no real or artificial intelligence will need to determine the metadata for your dental bills; instead, the software that generates the e-bill will embed the metadata (including that it is a dental bill, who it is from, the total, and other such information).

While some people speculate that we keep too much, we are actually frustrated that cost or copyright gets in the way of keeping absolutely everything that could be useful. The lack of an electronic version of every book we read is possibly MyLifeBits' greatest weakness. This is not because we want to read the books using a computer screen. Rather, we want the computer to "read" the book and help us recall things in it. In principle, we could have scanned our books, but since scanning costs are declining and there is some likelihood for their future availability, the decision was to not bother.

We have observed the more that is captured, the more correlation is possible to help find things. For example, suppose you want to refer to a document but cannot think of anything about it—but you recall viewing it when visiting Boston last year. A GPS trail entry can then be selected, and a search performed for all events from the same day, where the entry for editing the document along with its name appears. We could multiply examples of this sort—perhaps you recall it was a hot day; perhaps you remember an appointment on your calendar; perhaps you recall having a lot of windows open on your desktop. The more the system logs, the better the chance of having the "memory hook" that will help you find what you seek. We never regret capturing; but we often regret not capturing more. Storage space is essentially free and we can always add software to filter out less interesting items.

Some of the actual queries resulting from storing everything and being able to pivot or correlate using various metadata attributes include: finding the title of a book from an email message, invoice, or recipient's thank you; retrieving Web pages for a reference while authoring a paper and commuting on the train; finding a particular tile model that was used during a past home renovation by retrieving the contractor's specifications and invoice; recalling a distant colleague by looking at all correspondence about "storage"; replaying a stored phone message for a name or possi-

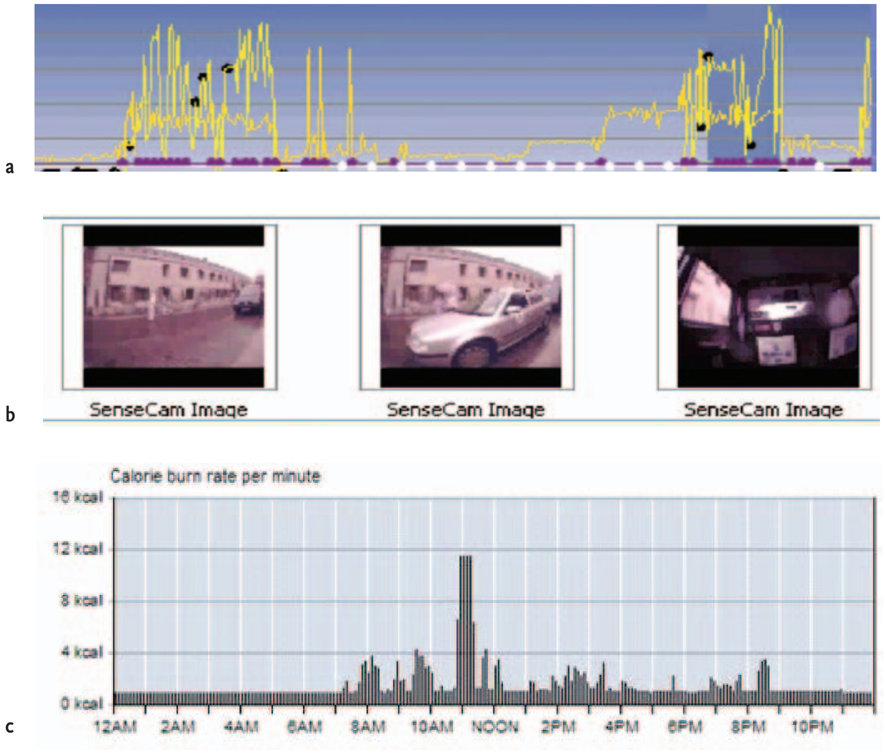


Figure 3a. SenseCam light intensity and vibration versus time.
Figure 3b. SenseCam photos taken based on triggering sensors.
Figure 3c. BodyMedia calories burned per minute using BodyBugg armband sensors.

bly sharing such a message; using a caller name to identify a particular call time to retrieve a Web page being viewed at that time.

While recall is critical, the collection is so large that the user cannot remember much of the contents, and will never search for them—in effect never "use" them. Thus, a killer app is the screen saver. Ours shows both photos and short video clips (selected from longer video files). The MyLifeBits screensaver allows us to enjoy pictures and videos much more while pleasantly refreshing our memories; videos were almost never watched prior to using the screensaver. Furthermore, the screensaver has been a great place to encourage comments and ratings. In the context of a family room, commenting on media has become a fun activity. The children join in, wondering: what will come up next? Who can say something interesting about it? Furthermore, we observe that the screensaver in the family room regularly elicits comments in the form of ordinary conversation; by capturing these comments their number and value increase.

With the vast flow of content including email, Web page visits, meetings, and so forth, coupled with the fact that we have powerful ways to search for content, one might conclude that no organization is needed. In effect, everything can be in one, large folder and items are retrieved by their content. This is the exact opposite of how the project started five years ago—over 30,000 items were named and placed in approximately 1,500 file folders, and retrieval was principally by name. Both views are valid—such organizational principles are the domain of classifications and ontologies including the Semantic Web. However, with large quantities of information, users are not just unwilling to do this, but are in fact unable to do it. Special skills are required to construct useful classifications. The first impulse we and others have felt is to just wish away the usefulness of metadata and hierarchies. But full-text search is not enough; in our experience, many items require some other attributes to be found. Furthermore, hierarchical organizational schemes have been developed with good reason: flat tagging systems have difficulty coping with scale. Hierarchy allows for broadening and narrowing one's scope in a meaningful way. To avoid having to become professional curators constructing our own personal classifications, we have become interested in classification sharing. We are experimenting with hierarchical classifications that will be developed by others to be downloaded by the user, and which contain extra information such as synonyms and descriptions to ease their use. One such classification we have developed is document type, which contains several hundred unique entries such as article, bill, will, business card, report card, greeting card, and birth certificate. Document type can be broken into a few different dimensions such as size, form, content, and supplier to enhance retrieval.

But even with convenient classifications and labels ready to apply, we are still asking the user to become a filing clerk—manually annotating every document, email message, photo, or conversation. We have worked on improving the tools, and to a degree they work, but to provide higher coverage of the collection more must be done automatically. The first, easy step is to stop throwing out any potentially useful metadata. Time is probably the most important attribute in our database, yet some photo-editing programs erase the value for the date the image was taken. Just having time and location would be a stride forward. Even capture itself must be more automatic on this scale so the user isn't forced to interrupt their normal life in order to become their own biographer.

One factor that discourages the use of new organizational techniques is the dependence of email clients, legacy file systems, and other applications on their own independent hierarchical structures. We agree with Boardman [2] that folder structures should be integrated—in our case, also integrated with our more flexible collection structure. Reporting tools with appropriate visualization are very useful applications. A simple query-based tool can be remarkably insightful and useful from "how I spend my time" to "count and space used" by different items. Reports can track what is being worked on or being thought about, for example by plotting the word "budget" or "nominating committee" against time. Figure 2c shows the mouse and keyboard activities on a hourly and daily basis for each active screen. In this fashion the amount of work on a document, spreadsheet, Web page, or other activities can be logged.

Programs that can assist in the creation or automatically create trip diaries and stories will considerably increase use, especially for future viewers who have no idea of the content. For example, a fishing trip diary with a timeline, animated maps, and annotations is substantially more valuable to us and our progeny than a collection of unlabeled photos in a labeled folder.

New capture devices vastly broaden the nature of personal recording. Passive picture taking using the sensor-enhanced (see Figure 3a) SenseCam is also very promising [5] whereby a camera captures several thousand photos a day (see Figure 3b) complete with voice comments, conversations, and location. Figure 3c provides still another glimpse of this future as a BodyMedia on-body armband logs every step taken, heart rate, and caloric output.

While we can foresee a time when everything can be captured, easily found, and utilized, it is not clear whether this capability will always be desired and in some cases allowed. For example, lifetime capture [4] raises many questions that lie beyond the scope of our research, touching on legal and societal issues (see the article "Digital Memories in an Era of Ubiquitous Computing and Abundant Storage" in this issue for more on this topic).

Conclusion

The first 50 years of computing were dominated by numbers and text. Most of the items in personal computers were correspondence including email, spreadsheets, papers, and presentations. The next era in computing is one in which PCs go beyond typewriters, calculators, and communication devices to capture, store, organize, and present a personal lifetime archive that expands to include multimedia (images, video, sound) and then goes even further. It is fundamentally a transaction processing system that records virtually everything in a person's life at meaningful resolution—a user's interaction with others, as well as logging location, calories, heart rate, temperature, steps taken, Web pages, mouse clicks, and heart beats.

No matter how many tools we add to this project, there always seems to be an inexhaustible backlog of new capabilities to add, and new questions to answer. We have, however, created a very useful database-oriented platform that can facilitate the exploration of these applications. MyLifeBits will serve as a platform for research as we continue to study the many issues related to personal lifetime storage. It has established a new benchmark for what we believe future personal computers must be. C

References

1. Bell, G. A personal digital store. *Commun. ACM 44*, 1 (Jan. 2001), 86–91.
2. Boardman, R. Workspaces that work: Towards unified personal information management. In *Proceedings of HCI2002, People and Computers XVI—Memorable Yet Invisible. Volume 2, 216–7*, London, 2002.
3. Bush, V. As we may think. *The Atlantic Monthly 176*, 1 (July 1945), 101–108.
4. Cheng, W., Golubchik, L. and Kay, D. Total recall: Are privacy changes inevitable? In *Proceedings of The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04)* (Oct. 15, 2004, New York), 86–92.
5. Gemmell, J., Williams, L., Wood, K., Bell, G., and Lueder, R. Passive capture and ensuing issues for a personal lifetime store. In *Proceedings of The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04)*, (Oct. 15, 2004, New York), 48–55.

Jim Gemmell (jgemmell@microsoft.com) is a researcher with Microsoft Research in Redmond, WA.
Gordon Bell (gbell@microsoft.com) is a senior researcher with Microsoft Research in San Francisco, CA.
Roger Lueder (rlueder@microsoft.com) is a software development engineer with Microsoft in Redmond, WA.