MyLifeBits: A Personal Database for Everything

Jim Gemmell, Gordon Bell, and Roger Lueder

Feb. 20, 2006

MSR-TR-2006-23

Microsoft Bay Area Research Center

San Francisco, CA, 94105

MyLifeBits: A Personal Database for Everything

Jim Gemmell, Gordon Bell, and Roger Lueder

Microsoft Bay Area Research Center

San Francisco, CA

Abstract

MyLifeBits is a system that began in 2001 to explore the use of SQL to store all personal information found in PCs. The system initially focused on capturing and storing scanned and encoded archival material e.g. articles, books, music, photos, and video as well as everything born digital e.g. office documents, email, digital photos. It evolved to have a goal of storing everything that could be captured. The later included web pages, phone calls, meetings, room conversations, keystrokes and mouse clicks for every active screen or document, and all the 1-2 thousand photos that SenseCam captures every day. In 2006 the software platform is used for research including real time data collection, advanced SenseCams, and particular applications e.g. health and wellness.

This article expands on the January 2006, CACM publication of the same name. MyLifeBits features, functions, and use experience are given in the main body, followed by an appendix of future research and product needs that the research has identified.

Introduction

The January 2001 CACM publication [1] of "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." 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 a moderately active professional person. This article expands on the January 2006, CACM publication [7] that 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.

Table 1. Bell's content (August 2005)

Item type	Number	Size (GB)
eMail messages	91,266	0.3
Web pages	67,491	5.4
Pictures	41,908	9.0
Doc & Rtf	13,445	1.2
Other	5873	0.9
Audio	5067	12.5
PDF	3215	5.0
Tiff	2821	8.1
PowerPoint	1772	4.6
Video	1301	62.7
Total	**234,159**	**109.6**

Note: TIFF and PDF hold about 250,000 pages as single and multiple document files.

Since 2000, 40 GByte disks at $10/GByte have been replaced by 500 GByte disks at less than $1/GByte, 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 around 0.5 GB/month, but in a non-linear fashion. Emails grew larger as attachments became more common, and digital photos took more space each time a new camera with more mega-pixels 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 1 GByte/month for the duration of an 80 year life assuming only modest storage for what is seen and heard (see Table 1). However, changing user patterns or technologies 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 [2]), or storing all meetings and conversations, or storing photos in raw rather compressed form. Inclusion of video can easily exceed one GB (GigaBytes) per day. Indeed, in a brief experiment recording television programs that might be watched we quickly acquired nearly 2 Terabytes. We now believe that a terabyte will hold a text-audio lifetime at 20th century resolutions and quantities, but speculate that 21st century users may expect to record their life more extensively and in higher fidelity – and may drive a market for much greater storage. Table 2 gives an estimate of storage requirement from various data sources on a daily, monthly, and lifetime basis.

The original project organized the corpus using careful file 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 we wanted to test different ways to organize and classify information.

Table 2. Daily items captured resulting in monthly and 83 year lifetime storage requirements

Item	Daily number	Month total (MB)	83 year Life total (GB)
1 MB Books|reports	0.1	3	3
5Kbyte Emails	100	12	12
100 KB Image scans	5	12	12
75 KB Web pages|docs	100	225	225
100 MB Music (1, compressed CD)	0.1	250	250
1 KB/s Listened audio (low quality)	40,000	1,000	1,000
1 MB Photos (Medium quality)	10	250	250
SenseCam photos (50KB)	1,000	1,250	1,250
2 GB/hr TV (S-VHS quality)	4	200,000	200,000

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, annotate, and utilize content. Also, we wanted to obtain a unified database in contrast to the many data "islands" being created including mail, contacts, and meetings, finances, health records, photos, etc. 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 sixty year old article is a prophetic blueprint that includes the computer, low cost library storage occupying a tiny space, commerce with automatic inventory control and billing, fast communication, speech interfaces, and a hypertext-links. Our interest is in his all inclusive, personal information system, which he called *"memex"*.

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

Bush posited 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." 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."

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.

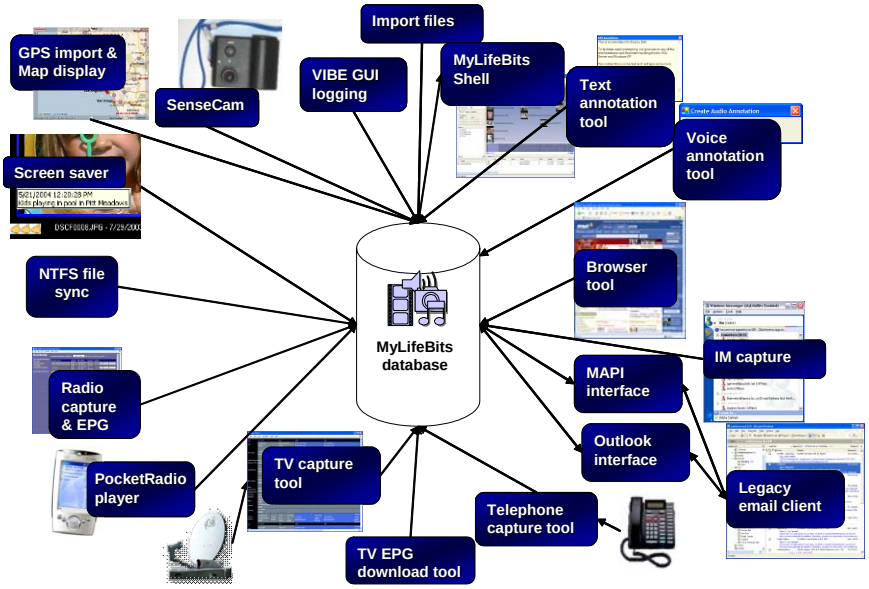

Figure 1. The MyLifeBits store and capture/display tools.

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, songs, etc., 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 that metadata are often a key part of user recall, e.g., that an email 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.

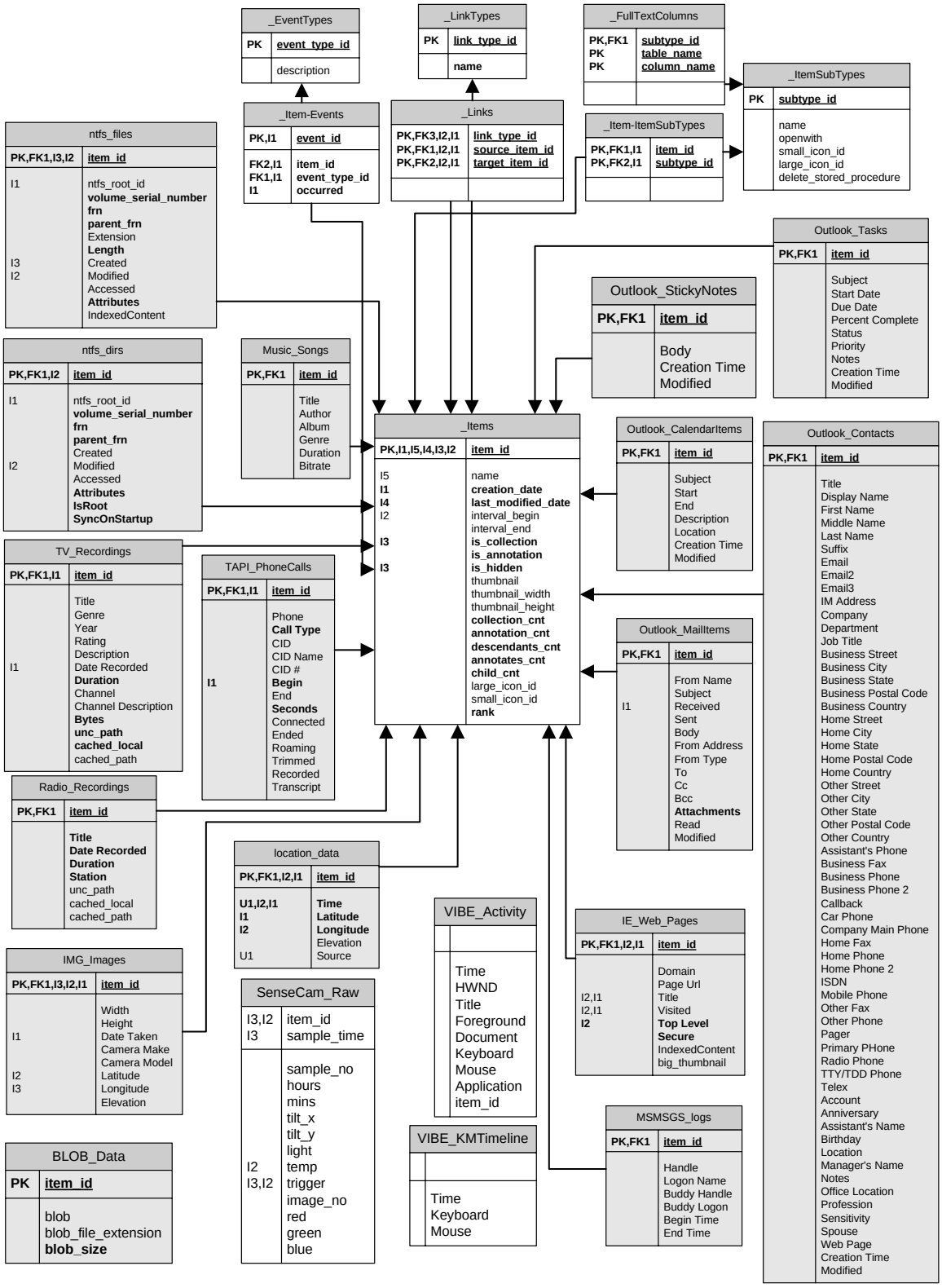

Figure 2. Key components of MyLifeBits Schema. Tables for type-specific metadata are shown in grey.

Once everything was in a database, the MyLifeBits project became a quest for useful tools to organize, associate metadata, access and report about the information. Figure 1 shows the many different capture and display tools used to populate the store and then to search or access it. 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 (IM) chat sessions, all telephone conversations, as well as meetings, radio, and television program usage as shown in the diagram. A GUI logger records all mouse and keyboard activity (see Figure 5). 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. It enables refinement or pivoting according to metadata and links as described in the following section. It provides for the creation of text and voice comments. For example, any number of selected items may be commented on using the annotation function with 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 slides shows with audio.

MyLifeBits has at its heart a SQL Server database that can store content and metadata for a variety of types, including contacts, documents, email, events, photos, songs, and video. Figure 2 illustrates the MyLifeBits schema. All captured entities that are exposed to the user are called "Items." Each item has about 20 common attributes, registered in the _Items table. Items are of one or more types: the _Item-ItemSubtype table indicates the specific the type(s) of each item. _ItemSubTypes enumerates the valid types. For each item type (such as email message or image) there is an additional table for type-specific metadata (grey boxes in Figure 2). For example, contacts have an additional 62 attributes, including email address and birthday. Currently, our database supports 25 item types. Table 3 enumerates the main tables of Figure 1 and lists some of the key metadata stored for different item types.

Links (_Links table) are a generic way to connect items. They connect source and target items and have a type. Link types are enumerated in the _LinkTypes table. Events for items (e.g. opened, played, paused) are tracked in the _Item-Event table. Valid event types are enumerated in the _EventTypes table.

Table 3. Tables for various aspects of the overall schema.

Table	**Key meta-data in schema**
Every item	ID, name, time, image, annotation, collection, descendant
Links	Mechanism for annotation, containers or collections, facets, photo-contact link
NTFS file, Legacy app	Location, dates, extensions, indexed content
Images	dimensions, date, camera, location (latitude-longitutde-elevation)
Music (song)	title, author, album, genre, duration, bit rate,
IE Web pages	Domain, page URL, title, visited, to level, indexed content
Outlook (4)	Calendar (7), contact (60 items), message (14), Task (10), Note (3)
Video cliplet	media start, stop, record begin and end.
Phone call log	time, call type, caller ID name and number, transcript,
TV record.	title, genre, year, rating, description, date, channel, size…
Radio record.	Title, date, duration, station
SenseCam log	time, tilt, light, temp, image no., R-G-B, trigger,
GPS log	locations
MSN iM log	Handle, logon name, buddy handle, buddy logon, begin and end times
Vibe logs	Time, title, foreground, document, keyboard, mouse, app,

All text fields in all tables are searchable using the SQL Server built-in text search (MSSearch) that handles stemming, similarity, and national languages. The IndexedContent column of the ntfs_files table is filled with a text-only version of the file contents (as extracted by an MSSearch iFilter), and full text search is then performed on this column. Likewise, the IndexedContent column of IE_WebPages contains a text only

version of the captured web page for full-text search, and the Body column in Outlook_MailItems and OutlookStickyNotes are extracted as text for indexing. Files are replicated to the database in the BLOB_Data table and the MSSearch iFilter for the given file extension is used to index the contents of files.

In addition to these core tables MyLifeBits, many of the MyLifeBits tools create additional tables for their own use. For instance, Figure 2 shows two tables generated by the PersonalVIBE GUI logger (VIBE_Activity and VIBE_KMTimeline). The MyLifeBits user interface, the "shell", has a number of tables, e.g., for saving queries, as does the program that interfaces with Outlook. In all, there are around 40 additional tables today.

SenseCam is a wearable camera and sensor pack that continuously records environmental information [2]. SenseCam data are imported into the SenseCam_Raw table. SenseCam environmental samples are items connected to associated photos using a link. GPS readings are stored in the location_data table. We debated whether they should be items, as the notion of annotating a GPS reading holds some interest.

Items can be linked together implicitly using time to "tie" them together as Bush suggested or by geographic proximity; or explicitly linked with typed links e.g., 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 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 – this is much more flexible and convenient for organizing information than folder hierarchies.

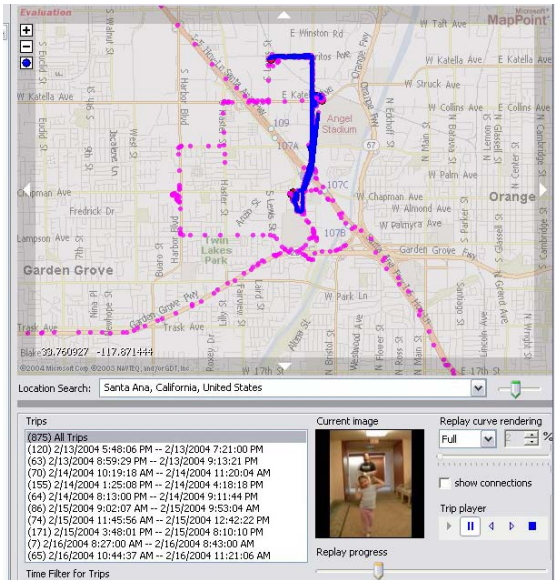

Figure 3 – Map interface: pink dots for GPS points, red dots for photos, and a blue line used in animated trip replay.

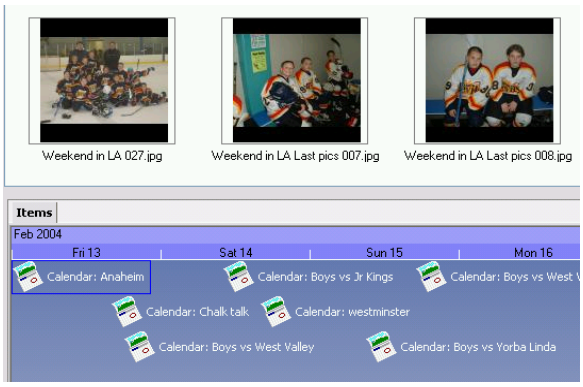

Figure 4 – A calendar becomes a photo-diary when photos are linked to events.

Metadata and linking are nicely illustrated by the way they are used to enhance 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, by time when taken or place where taken, or by subject, etc. Eventually, we expect software to analyze a photo's content to create additional metadata. At present, our software allows photos to be geo-located by dragging and dropping them onto a map to "link" a location with a photo. Alternatively, a common timestamp implicitly "ties" the time a photo was taken with a person's GPS recorder location to create location metadata. Figure 3 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 4. Similarly, a "person in photo" link can be used to manually connect a photo and contact.

Experience and Observations

Having a surrogate memory creates a freeing, uplifting, and secure feeling – similar to having an assistant with a perfect memory. Our only store is electronic, including financial and legal documents such as bills,

contracts, pay stubs, trusts, and wills. Stock certificates, paper money, and certain other legal documents are the only retained paper because they represent money or have special legal status.

Part of feeling secure is knowing that capture is automatic. While browser web capture at first struck us as somewhat trivial, this essential feature has changed our behavior. The failure of Gemmell's hard drive resulted in losing four months of captured web pages and was an emotional and productivity 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 that it was lost. Bell had a similar experience when he inadvertently deleted an email folder that while backed up was not properly synchronized (see versioning section of the appendix) and experienced a similar effect. 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 specs. As many internal sites are changing and transitory, having a personal copy is essential.

The good news is that more and more content is being "born digital" without the need for scanning. We expect that, over time, more and more 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 and many of these are available digitally now. RSS feeds from professional organizations will improve this situation. This will not only reduce the scanning effort; there is also the opportunity for metadata 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, procedure, time, etc.).

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. As we can't predict when some item (e.g., an old bill, conference announcement page, attendee list, business card, or scribbled post-it note) will be required, the easiest and safest thing is to simply keep it all. Everything. Many people have been taught to process and file incoming items – with Memex, the only option is to process or ignore incoming items, leaving an item such as an email that crosses our screen intact and potentially referenceable at another time and context. We may find a crazy idea from a friend 5 years ago, will turn out to be the key to solving a perplexing problem. The lack of an electronic version of every book we read is a weakness of MyLifeBits. This is not because we advocate, want to, or actually read books using a computer screen. Rather, we want the computer to "read" the book and in order to 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 in future for digital copies to be available commercially, the decision was to not bother. Although in the case of Bell's books, all have been scanned and are accessible.

We have observed that the more that is captured, the more correlations are possible to help find things. For example, suppose you want to refer to a document you recall viewing when visiting Boston last year. A GPS trail or travel calendar entry can be used a the starting point for a search performed for all events from the same day, where the entry for viewing the document along with its name and thumbnail 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" which 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 irrelevant 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, 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 in a 5 year old home renovation by retrieving the contractor's specifications and invoice; recalling a distant colleague by looking at all correspondences about "storage"; replaying a stored phone message for a name or possibly 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 search/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 vastly increase.

Organizing by "lives" and time

With the vast flow of content including email, web page visits, meetings, etc. and 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 with no attention to "location". This is the exact opposite of how the project started five years ago – over 30,000 items were named and placed in about 1500 file folders, and retrieval was principally by folder location and file name.

The classifying principle we used in the original folders is still valid and useful. To break the problem into more manageable parts, content was divided in four ways, using two orthogonal attributes: time and "lives". At the topmost level items are classified as either belonging to one's personal life or professional life. Below that, they are divided by age into either current or archive. Archive handles items associated with a past event, or some period that can be treated as read only. E.g. college years that that had already passed and relationships with a former company or organization. Any ongoing "life" or professional activity is held in the Current subfolder. Personal "lives" were, generally not migrated into archives because of their on-going nature although nearly all photos were treated as archival. The rationale was that photos are fundamentally archives and this had to do with the mindset that Windows established by making "My Pictures" a special folder, that would lead people to store all photos in one folder hierarchy. Most of Bell's photos are in a folder called "A Big Shoebox". Over time, however, we also ended up storing photos throughout the folder hierarchy because all items pertaining to a single subject, e.g. letters, articles, web pages, photos, or whatever are most likely to be held together, and separating photos into another folder hierarchy based on their data-type, results can result in duplicating folders. Photos represent a particularly nasty problem because they inherently require a database/link approach for filing: who, where, when, what, etc. Photos are the archetypical example of why hierarchical folders fail and the need for metadata, links and text-annotation. This has given rise to the dozens of photo database applications.

Both a classified content and unstructured data view are valid and necessary. Organizational principles are the domain of librarians creating classifications and ontologies including the semantic web versus machine learning approaches to self-organize content. However, with large quantities of information, users are not just unwilling to classify, 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 lets you broaden and narrow 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, 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, 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 date taken. Just having time and location, would be a stride forward. Even capture itself must be more automatic on this scale so that 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 [Boardman02] that folder structures should be integrated – in our case, also integrated with our more flexible collection structure.

Reporting tools with appropriate visualization are another class of very useful apps. 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 e.g. by plotting the word "budget" or "nominating committee" against time. Figure 5 shows the mouse and keyboard activities on a hourly and daily basis for each active screen taken from the PersonalVibe GUI log. In this fashion the amount of work on a document, spreadsheet, web page or whatever can be logged.

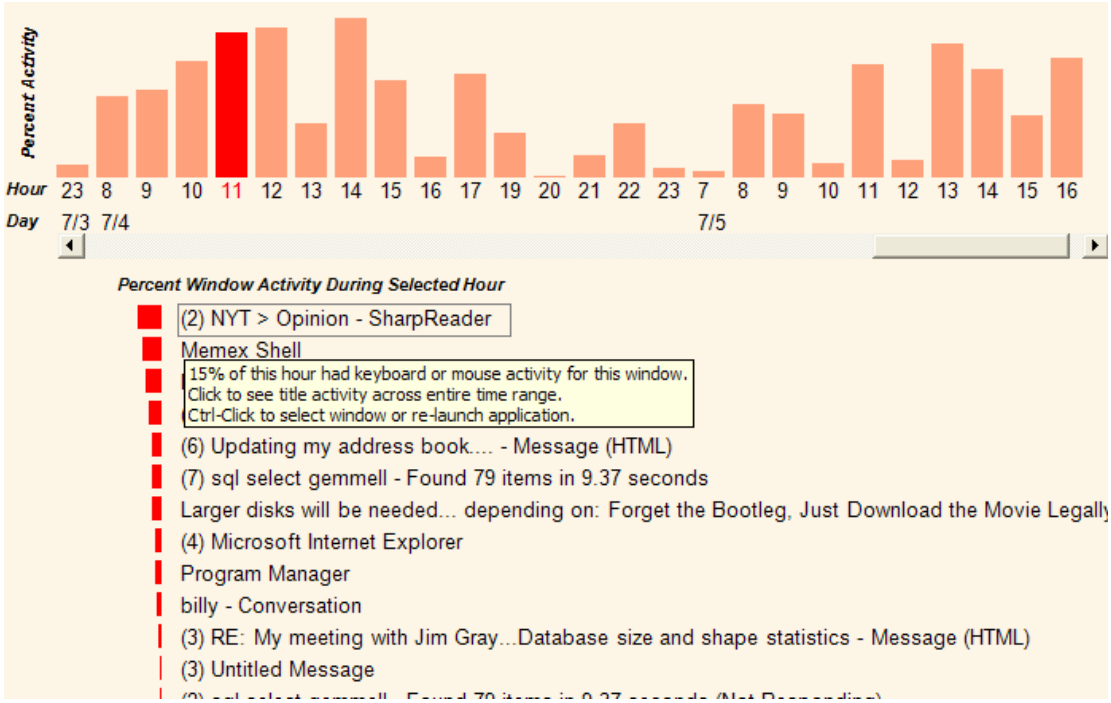

Figure 5 – GUI Activity log on an hourly or daily basis from the PersonalVibe GUI log

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 unlabelled photos in a labeled folder.

New capture devices vastly broaden the nature of personal recording. Passive picture taking using the sensor-enhanced (Figure 6a) SenseCam is also very promising [2] whereby a camera captures several thousand photos a day (Figure 6b) complete with voice comments, conversations, and location. Figure 6c 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[5] raises many questions that lie beyond the scope of our research, touching on legal and societal issues as described in Digital Memories [8].

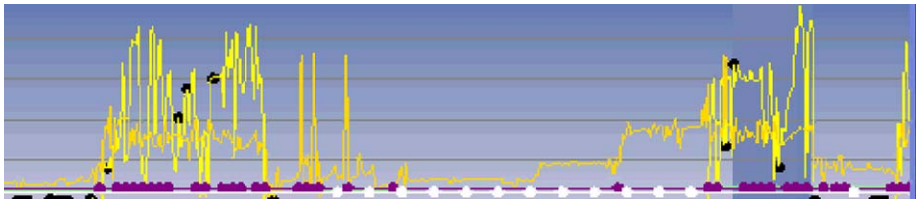

Figure 6a. SenseCam light intensity and vibration versus time.

Figure 6b. SenseCam photos taken based on triggering sensors

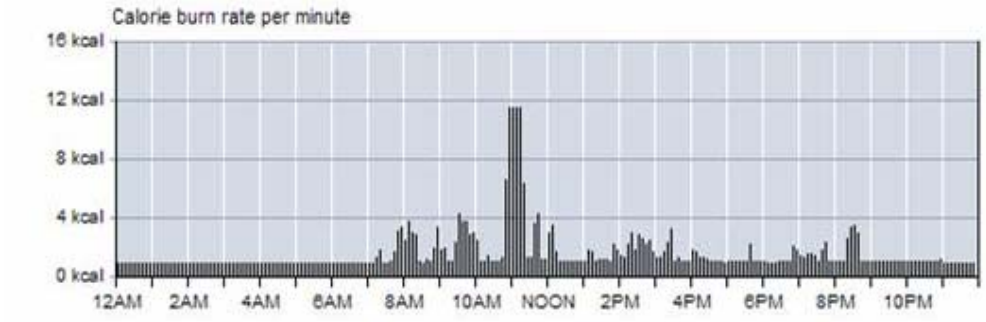

Figure 6c. BodyMedia calories burned per minute using BodyBugg Armband sensors.

Conclusion

The first fifty 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 especially as new data arrive e.g. health records. We have, however, created a very useful database-oriented platform that can facilitate the exploration and creation 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 set a new mark for what we believe future personal computers have to be.

Acknowledgements

The authors would like to thank all of the people contributed ideas, programs and encouragement. Jim Gray provided much of the inspiration and consulting on the database aspects. Kentaro Toyama and the world wide media exchange (WWMX) provided the basis for mapping photos. Mary Czerwinksi, Brian Meyers, George Robertson and Greg Smith provided the GUI logger and visualizations. Lyndsay Williams and Ken Wood gave the impetus for CARPE (Continuous Archiving and retrieval of Personal Experience) with their work on SenseCam. Astro Teller, Founder and CEO of BodyMedia provided the tools to let us

envision the importance of continuous body monitoring. Marc Davis and Larry Rowe of UC Berkeley gave us invaluable advice. We were also helped out by some excellent interns: Aleks Aris, Joshua Blumenstock, Evan Salomon, and Zhe Wang.

References

[1] Bell, Gordon, A Personal Digital Store, Communications of the ACM, Vol. 44, No. 1, January 2001, p86-91.

[2] Gemmell, Jim, Williams, Lyndsay, Wood, Ken, Bell, Gordon and Lueder, Roger, Passive Capture and Ensuing Issues for a Personal Lifetime Store, Proceedings of The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04), Oct. 15, 2004, New York, NY, USA, pp. 48-55.

[3] Bush, Vannevar, As We May Think, The Atlantic Monthly, 176(1), July 1945, 101-108.

[4] Boardman, R., Workspaces that work: towards unified personal information management, presented at the HCI2002 Doctoral Consortium, in Proceedings of HCI2002, People and Computers XVI - Memorable yet Invisible, Volume 2, 216-7, London, 2002

[5] William Cheng, Leana Golubchik, David Kay, Total Recall: Are Privacy Changes Inevitable? a position paper, Proceedings of The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE '04), Oct. 15, 2004, New York, NY, USA, pp. 86-92.

[6] Bell, Gordon "Dear Appy, How Committed are you? Signed Lost and Forgotten Data", ACM Ubiquity, 21 February 2000, Issue 1. http://www.acm.org/ubiquity/views/g_bell_1.html

[7] Gemmell, Jim, Bell, Gordon and Lueder, Roger, MyLifeBits: a personal database for everything, Communications of the ACM, vol. 49, Issue 1 (Jan 2006), pp. 88.95.

[8] Digital Memories (Memex) Request For Proposals (RFP), Microsoft Research http://research.microsoft.com/ur/us/rfps/RFPs/DigitalMemories_Memex_RFP.aspx

Appendix: Future Research & Product Development

MyLifeBits is a platform for applications that uses a database with streams and transactions. It deals with the acquisition of new data-types (e.g. web pages, phone calls, user activity, and SenseCam streams) that quite often require special applications, interfaces and visualization to be effective. By 2006 we already have three different shells, and have had to make each readily extensible so that new visualizations can be plugged in.

To realize the memex vision for everyone certainly requires more fundamental research and technology, as well as understanding whether anyone really wants to recall their bits. In fact, we continually hear claims for the need to forget because that's supposedly how people cope with unpleasant events or periods, or their lives in general. They simply don't want to be reminded of things past.

In this section, we outline issues that have arisen through use, questions about a memex future, and just plain additional features. The "Dear Appy" problem inevitably comes up as number one – how do you insure that your bits will live forever and be interpretable? A close second is how does the system know what was said, or identify who or what's in a picture or segment movie scenes? Information control is perhaps the most threatening to individuals: what is in my memex that I or others own or can freely use, or no one can access. What's needed is being able to control and protect content, including insuring that items are not discoverable or even able to be forgotten, for sure – this being very difficult with good backup procedures, etc. The multi-persona problem is real: memex has a single data store for everything personal and professional across many organizations. In effect, we tend to think of ourselves as having multiple lives. When a person leaves most commercial organizations, their files must remain – again, this implies Memex needs the ability to absolutely forget. This is increasingly difficult with links and annotation. With everything becoming digital, where does all of life's bits reside? And can all or at least the relevant bits of one's life be collected? Can it this be automatic?

The "Dear Appy" Problem: Insuring that a memex lives forever

"Cyberdata" appear to be no more at risk than physical "bits" as we have seen when natural disasters e.g. earthquakes, fires, hurricanes occur. In fact, for valuable content that is in every day use, there's no reason to not have backups that will survive disasters. Physical media is at substantially higher risk because it is very expensive to duplicate! Of course, like any file or database, it must be backed up to prevent loss of data. Recently, a large email folder was deleted because it was not properly moved to a store accessible to the MyLifeBits store before it was deleted from the corporate exchange server. The oversight wasn't detected for 3 months. This merely pointed out the necessity for periodic snapshots. Each of us has only one lost data incident that a more elaborate backup system could have avoided.

However, a critical problem in keeping digital content is the ability to use a given format in the future. In 2000, Bell wrote an article titled "Dear Appy, How committed are you? Signed, Lost and Forgotten Data" that began to explore the problem [6]. At the time, the problem seemed straight-forward and perhaps solvable: we just needed a system in which stored content would "live forever" (i.e. be playable/viewable) via conversion to a few, simple "golden" standards e.g. text, rtf, tif, jpeg, or mpeg. These formats were expected to be supported for "nearly forever" – or at least be capable of migration to each new generation. Content from a database such as Outlook or Quicken would be extracted or flattened as reports coming from labeled transactions and put into files. This requires work – every, say, decade the files are converted to a more modern format. This is similar to taking care of cemeteries.

So in 2001, our concerns were media migration, hardware and operating system migration, and having a few applications that would always be able to render content stored as a small number of data-types versus the need to support a unique program. Five years later, having built memex, a unique platform with a variety of one-of applications makes the problem of storing data forever much more difficult. One solution is that you just extract the data like we posited in 2001, and migrate the data every decade. Of course, this means no dynamic rendering or interactivity as with an application program.

The best and perhaps only real solution is a commitment to migrate the platform and its data forward every decade or two such that the system is emulated on a more modern system. As long as technology continues to evolve, then running an old app and storing old data is trivial on the latest computer e.g. c1995 PC disk drives were 100 Megabytes, and 10 years later, 100 Gigabytes, so the content is negligible. Fortunately, there is work on such emulation systems. However, rendering data statically using xml tags to describe the metadata may be a practical solution when considering the evolution and variety of systems such as Memex.

Clearly, needing a Memex with all the applications and legacy data that would provide interactive access, can only be provided by a working Memex!

Automatic Recognition of speakers, speech, sound, photos, and video

Since most searches rely on text as for retrieval, speech to text is greatly desired. As we acquire more content associated with voice annotation, notes, telephone calls, meetings, conversation, room environments, SenseCam audio clips, voice note recorders, and video the need increases. Each of these situations are highly idiosyncratic for which some technique can be used to enhance the value of the content, even it high quality speech-to-text doesn't work.

SenseCam II and many commercial cameras allow some audio to be recorded and associate with a photo. One use of this would be when greeting a new person: the speech at the time would include the person's name, and ideally the name of the person would be able to be associated with the photo and then put into a contact. Depending on the value of the situation, one can imagine using a few recognizable keywords to signal that audio transcription is needed. In the case of the SenseCam this might serve for personal introductions, notes, and scene annotation. SenseCam and any other device including video monitoring equipment that records multimedia content are enabling new classes of applications such as those described below.

Overall, there are a myriad of situations where signal analysis of sound works quite well and can be incorporated. We have not found a reasonable way to analyze phone messages, for either individual speakers or in general and hence knowing the time and location of the call is the only useful meta-data.

Speech-to-text dictation, including content from voice recorders works extremely well! It is unclear whether phones or SenseCams will be able to achieve a useful state. In addition to speech analysis, sounds could be categorized, or used to detect location (e.g. office, train, or restaurant) to further describe one's life.

Photos have a similar problem that requires image understanding – who, what, where, when identity. The immediate solution is to accurately label each photo with time and place. Location is imminent with new cell phones and knowing location will eventually be part of cameras. Ideally, however, content analysis i.e. who and what are needed. With 2005 image analysis technology, useful meta-data can be applied to identify scene types and similar images. Individual person recognition with manual verification could be useful in a few, limited contexts such as identifying named family or group members. However, in 2006, the error rate is too high to be generally useful.

Video, of course, contains audio and images, and hence has all the same problems as those media types. Additionally, it is desirable to be able to break it into segments, meaningfully summarize it, and to understand actions (i.e. image understanding with a time dimension added).

Digitizing everything, everywhere, all the time --Scaling in space, cyberspace, and time

The conversion of all information to digital form, and the high value of a complete store with many data types in great quantity will create a compelling motivation to access everything, everywhere. We see a progression from personal computer systems like we envision for Memex, to organizational Memexes whereby every event is captured throughout a company, government agency, professional meeting, or whatever. Privacy, ownership, rights, etc. will be the key impediments and yet safeguards for this vast content. These are discussed in the next section.

Going beyond a PC to a distributed environment is a change in our design philosophy whereby everything is in a singe machine as envisioned by Bush. We have assumed that anything accessible outside a person's Memex such as a web page is transient[2], and hence a copy must be created in one's own Memex. The network cloud with "IP on Everything" inevitably connects most all of the information sources one would like to capture. A simple solution would be for agents in every system to send items back to a person's Memex. For practical reasons, a server might be their Memex. This introduces the notion that one might have a partitioned system, much as we think of MyLifeBits as a series of lives with content being distributed on an ownership basis.

Personal capture devices including cell phone calls and conversations, camera photos and videos, SenseCam images and data would all be transferred back to everyone's personal Memex. Home, personal and entertainment items e.g. books read, interactive games played with fellow contestants, songs plays, and videos watched would be transferred in this fashion. Within the next few decades, objects in a home will need to be queried or want to communicate their states --- container and item content and location, health and home environment monitoring, etc. Active RFID tags will go beyond communicating an item's name to providing state information e.g. a carton of milk's age and fullness. Schema are ever increasing with smart devices that create new needs.

In professional and working environments we envision room and office conversations, and meetings would be captured. These would be part of a person's corporate persona.

No doubt content of this type which is "owned" by an organization would be kept at a central site… and only rarely would be allowed to become part of one's persona.

With more content, the size will no doubt exceed our Terabyte –we have no idea about the size, whether we can maintain the design that puts everything into a single system, and the scalability of a Memex. Clearly, as in the case of video, it is impractical and even unnecessary to save content in every person's Memex. We

[2] For instance, we do not trust a vendor to to retain an electronic invoice for us – they may go out of business, or tamper with the record is there is a dispute.

know that we'll have to rely on URLs for size practicality, ownership, and security. Certain parts of Memex need to and will be distributed.

Control of the bits: ownership, security, privacy as well as authenticity, deniability, fabricating, forging, and forgetting

Since a Memex is an evolution of current file and databases, it can be argued that it doesn't require unique research. However, as we evolve to record everything for all time, there are significantly more opportunities for the invasion of privacy. Our arguments for Memex have been as a surrogate memory for the brain and as such, we may need all the control mechanisms that an individual uses: intentional forgetting when deposed, adding noise or creating misleading information, and just plain removal of information for all time. Many of the issues relating to control involving ownership, privacy and security, and modification rights of Memex bits are being defined and redefined with new laws in various countries. Authenticity is really an important factor in a Memex, especially as one's Memex might be examined by one's progeny or biographers.

It is illegal in some countries to merely make a copy of anything, including an owner's personal items such as articles, books, and CDs that are protected by copyrights. Similarly, photos of people or objects are protected by copyright such that it is necessary to have permissions for these images. Of course, this creates the opportunity for a legal feeding frenzy around the content that a SenseCam captures. Some of these issues may already be clear based on camera, video camera and phone conversation capture – for example, in some US states such as California, phone capture is permitted as long as the party being recorded is informed. MyLifeBits dialed or received calls opens with a statement that the call is being recorded and notice is deemed equivalent to an opt-in. In other places, it is illegal to record. We are unaware of any place that permits unannounced or unsanctioned recording. Phone calls are really jointly owned, independent of permissions. Hence the ability to share jointly owned content makes the problem even more complex.

Undoubtedly the greatest concern when people hear about Memex is the ability to extract its contents for whatever purpose e.g. subpoenas, hence giving up one's private communications. Taken over a long period, every person has something in a Memex that is private – it might be considered embarrassing at some future time either taken in or out of context, is owned or co-owned by someone else, or be shared knowledge that would harm the individual or some other person or organization. Ideally, technology could be created to protect information that, prior to Memex, would have been kept in a private part of one's own brain. Clearly, over the long term, Memex has to be able to expunge information that is owned by another organization when a person is part of that organization. Research is clearly needed to deal with these issues that, in effect, separate an individual's separate lives.

Having a tight degree of control over the content by a Memex owner is essential if we consider our system to be a surrogate memory. This would allow one to be able to selectively publish content for various applications such as blogs, web sites, use by others. More importantly, if would clearly allow one to "forget". Of course, the flip side is fabrication of fact. Undoubtedly, a more important capability is the ability to insure that content is both authentic and unalterable. One may be tempted to "edit" historical documents, for example, to bring forecasts in line with really happened.

Versions and Organizational Relationships: Time Varying Views of Content, Family trees and Organization Charts

Most modern file systems don't support versions. One exception was the VMS file system of the 1980s that would keep a fixed number of past versions of a file. Most file systems today force users to save the file with a new name when they want to keep a version.

Versioning is needed for more than just files. For instance, many elements of a contact are updated over time such as address, organization, or phone number. Many times these are worth keeping a historical record in case retrieval by an old value is desired (e.g. I want to look up a past neighbor by our old street). In order to recover from mistaken updates, it may be worthwhile to keep old versions of meta-data and links. It may prove useful to even recall past versions of the database schema.

Evolution of existing schema and the creation of schema are critical in order to incorporate new devices and information types. We have developed a schema differencing tool, and a methodology for applying schema version changes. However, further research is required to make schema upgrades safe and easy.

Organizational Relationship are equally important for Memex. While links allow items to be related to one another, we require more tools to support this capability, e.g. to create organization charts and family trees. This is especially useful for "contacts" as used by Outlook and other mail systems. Given the incredible mobility of individuals in our personal and professional "contact" list, it is essential to have a contact for just a company and position, independent of any name that currently holds the positions. For example, bankers, brokers, doctors and other professionals that we maintain relationships with are much more transient than any of us would like. Updating this content is a continual hassle that Memex would ideally handle.

Memex Enabled Applications

One of our suppositions in building MyLifeBits with the SQL database was that various application islands could supply information such as a contact, physical location, organization just once and that this would not have to be "copied" to databases in other applications. For example, knowing the time and location of a photo can link to people, building and site names, etc. It remains to be seen whether this benefit can materialize because it requires each of the applications to utilize a common schema.

Will the user interfaces in the applications such as Money, Outlook, or Personal Health recording change once a single database is used, versus the use of existing applications specific databases?

Digital Memories Research

In June 2005, Microsoft issued a request for research proposals focused on Digital Memories and Bush's Memex vision [8]. Winners would receive funding, SenseCams, and the MyLifeBits software. 80 proposals were submitted and 14 accepted. The submissions covered a wide range of use and infrastructure from security, image analysis and classification, to education, memory enhancement for normal and memory impaired persons, tourism, and engineer's or scientist's notebook as envisioned by Bush.

In an applications environment such as education and health, the interest was alternatively on how people carry out functions e.g. learning, remembering, and also tools to recall lectures, tours, or daily life by helping Alzheimer's patient memory recall. One can imagine, in the case of helping mentally challenged persons, the user interface problem is challenging and will undoubtedly require some surveillance functionality with human assistance depending on the recall task. In 2005 there are is a great deal of R & D aimed at senior citizens in order to help them live longer, independently. Many of these projects include wireless sensor networks to make observations on a person's activities e.g. have they taken their medication, have they used an appliance or eaten, gone to the bathroom, have they fallen. A prototype bed has the ability to know whether the person is in bed and sleeping soundly as well as measuring HR, weight, etc.. Once an activity is sensed, a computer can provide reminders.

Storing Personal Health Information in Memex

A "Health Memex" (HM) would hold all of a person's health records, associated health financial transactions, and wellness information to aid in their understanding and management.

Personal Health Record (PHR). Health service provider reports including information performed for a person (e.g. exams, prescriptions, procedures, and test values) will eventually be available to the system in digital form using XML. For the next decade health transactions will unfortunately continue to be paper based. Changing this will require a system to hold the information e.g. HM or some dominant health records standardization efforts. The PHR has value for subsequent attending providers either when referrals occur, while traveling or when changing providers. Also, a person's complete record would be useful in large comparative studies. History coming from a complete record is key to understanding subsequent problems e.g. Type 2 diabetes that may arise.

Health financial transactions are likely to be the most useful aspect of HM. These consist of health events and chronic conditions that cause health providers to deliver a service e.g. diagnosis, procedure, prescription that are paid for by individuals and their insurers. This blizzard of paper, email, and phone calls to providers, and insurers includes explanation of benefits, provider bills, insurer records, and payments, together with errors arising in the transactions. Noise, in the form of errors and exceptions associated with payments, operate as amplifiers to increase flow. The principle user is the family financial manager who must deal with all the transaction from simple and compound event e.g. hospital procedures involving every type of service, and chronic care.

Wellness health records (WHR) consist of a person's vital signs and other metrics that are maintained for all time. These include: blood pressure, diet (including medication & vitamins), exercise (calories and heart rate versus time), sleep, gum depths, heart rate, blood pressure, weight, and chronic illness metrics like cholesterol. Genomic tests and ultimately a complete mapping will increasingly be a part of the record, enabling better understanding about one's future situation. These records form a first line indictor of health of an individual. They are control metrics that an individual and physician would look at to understand person's state. Changes in these metrics would flag potential changes in health. For example, one of us having had gout attacks is taking Allopurinol to lower uric acid – over the last few years, levels have increased that have not been noted by physicians, because the values lie within range.

The goal of Health Memex is to be able to delegate the caring and responsibility to individuals by making all of a person's health and wellness records, including diet, exercise, and environmental metrics known to themselves. Awareness coming from a closer understanding of the financial aspects of health care might allow competition. However, who of us has ever seen a price list of health care items?

The system should be responsible for maintaining and recording every event associated with a person's health and wellness. This means all interactions with the medical and wellness communities as well as their diets and activities. For wellness, this implies knowing where and when every calorie is consumed and spent – the system would know about diet and exercise, as well as weight, HR, and BP because these would be automatically captured and fed into the system.

We assume the average person has a minimum of one chronic malady that is being constantly monitored or treated e.g. asthma, diabetes, eyesight, glutton intolerance or allergy, heart, lung, or kidney disease. Maladies grow with time… and then you die. Most medical expenses are incurred in one's latter years.

BodyMedia's BodyBugg is quite useful in being able to report about sleep, waking, walking, sitting, and exercise, etc. It has the ability to detect sedimentary states e.g. being at a computer, TV watching, reading. In addition, it can detect degrees of stress from skin galvanic measurements. We expect other body monitoring devices to be forthcoming this decade.

Agent Technology: Being Actionable based on state

The health and wellness applications beg for and proactive Memex. Virtually none of the programs we use today outside of calendar are proactive – i.e. drive us to an action. If one looks at the whole area of elder care, the implication is clear – the computer has to take on the responsibility for care that might go beyond simple reminders and having an ability to inform some alternative care agent when certain alarm events occur.

Memex as a Service that would provide for community

What if a person's Memex in the cloud? While there are times when we want the Memex on another system for performance reasons, in certain situations, having Memex as a service may be necessary and desirable. For example, the problem of ownership of corporate data might better be solved by having a central system that holds my Memex for my life within a corporation. This can be implemented with a client application that would move content to a server when it is associated with "corporate life". Thus, one could use a single computer that would hold all of ones other lives e.g. family, professional associations, and the like that would be separated with activities within a corporation e.g. administrivia, corporate email, projects, strategy. In this way the ownership and rights would be clear.

Implementing Memex as a service would also have the advantage of allowing individuals to work together in a project fashion. A systems like this is apparently by implemented for intelligence analysts where community problem solving is important.

Enhancing PersonalVibe GUI logging to monitor everything

PersonalVibe GUI logging has the potential to provide much insight about how we use our computers. Right now it chronicles the activity details of each application, e.g., document edited, or web site visited. In the future, it should analyze activities for all applications including their degree of parallelism. For example, MSN Messenger or Communicator, Media Player, the telephone, and Outlook are used concurrently with all other apps. Similarly, several Office documents may be active at the same time and only when they are activated, is it clear they are part of a task. Knowing these use characteristics can be helpful in understanding and improving an individual's performance.

Cell phone use data can also be captured and analyzed.

Personal Computer Network traffic and activity represents another fruitful area for monitoring if it provides any new insight. For example, in Australia, one service allows unlimited amount of download data each month at 300 KBps until 10 GB is downloaded, at which point the channel is "shaved" or throttled down to 50 KBps. A user operating VOIP, video conferencing, or listening to internet radio typically requires about 100 Kbps, adding up to 360 MB in an 8 hour day. Thus 10 GB is good for 28 days --- something this user recently experienced.

Finally, GUI logging can give us overall insight about how we are spending time as we aggregate the total amount of time spent on a given document, the relative efficiency of actions on it, and in general where all the time goes for all applications. For administrative applications like travel reservation, it will be quite useful to understand the time various users take, and whether there is the expected learning curve for a new task. The only problem: users may not want to know, and certainly they are probably unwilling to have the organization that employs them know.