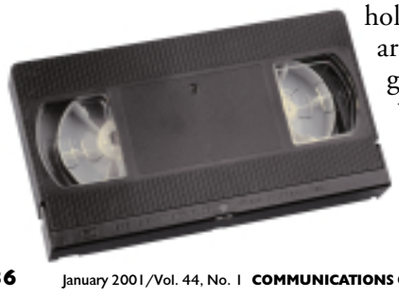
GORDON BELL

Finding a place to efficiently store all of one's digital materials.

A PERSONAL DIGITAL STORE

CyberAll[1] is a project to encode, store, and allow easy retrieval of all of a person's information for personal and professional use. The archive includes books, CDs, correspondence (such as letters, memos, and email), transactions, papers, photos and albums, and video. In 2000, only 16 gigabytes are required to store all the media in my personal and professional life—at a cost of $160 for disk storage. Two gigabytes are expected to be added next year. Encoding, indexing, and data-management costs far exceed the storage expense. The challenge is to automate capture, search, and retrieval.

CyberAll is a personal ontology [5] in contrast to a library [6] or Kahle's effort to archive the Web and television channels (see www.archive. org). It is my store for documents, photos of people and computing artifacts, music, and videos as described by Bush in [2] and Gates in [4].



CyberAll also holds reference articles, clipped graphs that heretofore would be physically stored, computer manuals, and magazines. At present, books are in "atomic" form; but CyberAll will include them as they become e-books.[2]

Within the next decade personal computers will be capable of storing a terabyte of information on an individual machine. In 2000, 40GB drives costing $400 are more than adequate to hold the content for most of a professional's lifetime reading, presentations, and audio recordings. A CD encoded at 128Kbps can be stored at a cost of $0.60. A user's CD collection is likely to use more storage space than the user's computer-generated and scanned paper files.

The next phase of CyberAll will capture conversations, interviews, meetings, and presentations. Recording speech from one's personal and professional lives will require over a terabyte (at 8Kbps)—but only a modest 25GB/year. Video is even more challenging. For home use, a terabyte holds 500 hours of DVD quality video and 1500 CDs, but more compression increases the capacity by a factor of at least 10. Recording a lifetime of everything seen requires 100TB. Doing this economically is still more than a decade away—it would currently cost more than $10,000 per year. But in two decades, it should cost only $100 per year and require an

[1] Cyberall and CYBERall are protected and copyrighted names of United Services International, Cyberall.com.

[2] www.research.microsoft.com/~gbell/CyberMuseumPubs.htm holds items of historical interest such as Hollerith's Patent, Amdahl's Law, various Digital Equipment Corporation documents, including those for PDP-1 and CDC 6600, manuals, posters, photos, and a talk about Seymour Cray.

infrastructure unlike anything we currently know.

The technologies for cyberization are improving at the rate of Moore's Law—doubling every 18 months. These include processor speed, storage capacity, scanner speed and accuracy, camera resolution and software, OCR accuracy and capability (for example, scan-to-HTML), audio and video encoding, printing and display, and standards. Thus, one can always wait for a better system or standards—things will be twice as good in 18 months. However, content and capture cost are almost acceptable, and the longer we wait, the more information is lost forever, so it is important to begin the process.

Based on my experience of being able to only go back 25 years for some content, the most serious concern of CyberAll is choosing formats that will be readable in 10 to 50 years. CyberAll requires a mechanism for carrying data forward from legacy media, systems, and programs.

Motivation and Goals

The motivation for CyberAll ranges from the technical challenges—because we can—to a desire to have an exhaustive archive. Electronic filing cabinets such as Ricoh's eCabinet [3] accept both computer-generated and scanned documents and index the documents they hold. Filing systems such as those used by Microsoft Windows 2000 and Office index their documents.

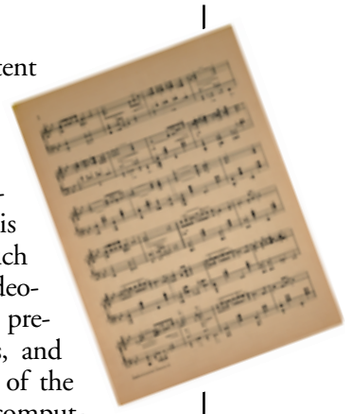
Many people have a "pack rat" mentality, and attempt to store everything possible to remind ourselves or others. CyberAll is an attic to store everything that can answer a question or explain what it was like when. It is a memory aid and a device to help tell stories. For some, this might mean storing everything—our first drawings, school report cards, and home videos. New Web sites such as www.123456789.net, www.legacy.com, and www.memorymountain.com offer to store letters, essays, photos, and stories "forever" and pass them on to their future generations of users.

Another goal of the CyberAll project is to understand the problems of coping with the exponential increase in the amount of information (for example, email, Web pages, images, audio, and video) that is becoming part of both our personal and professional lives. Given the tools to mass-produce documents, we are forced to become filing clerks!

The goal of CyberAll is both to encode everything and to eliminate paper that is used for storage (filing) and transmission. Paper will remain a dominant reading interface where its advantages are well known. Many documents that represent money—paper currency, bank notes, stock, and cancelled checks have to be retained.[3]

Using and Accessing CyberAll

Table 1 shows the kinds of content that occur in an individual's personal and professional lives for archival (mainly reference) and daily (working) use, such as contracts, email, and music. This includes encoded legacy content such as papers, photos, audio and videotapes to computer-created papers, presentations, photos, "ripped" CDs, and videotapes. CyberAll can play all of the content from photos to CDs on computers, home stereo, and TV sets.



CyberAll is for personal use as opposed to providing a general server. CyberAll operates in my COMOHO (commercial office, mobile office, and home office) environment, providing access anywhere, anytime. The main desktop computer in the BARC lab (CO) holds all files and is well backed up. The author's portable computer (MO) contains a large subset "cache" of the CO. It is the principal computer, used in the MOHO environments. In MO locations, modems, hotel LANs, and so forth communicate via the corporate network to CO for "uncached" documents.

In the HO, ADSL and cable modems link to CO, allowing audio and picture files to be "played." By keeping all information, CyberAll should be able to provide a useful set of answers and services including:

[3]The financial community—hiding behind "user resistance"—is decades behind in dealing with items as pure bits. The June 2000 law approving electronic signatures eliminates one more barrier.

Table 1. Data types and use for timeliness and user context.

User Context/ Timeliness	Personal (entertainment and personal finance-related)	Professional (work-related)
Archival (historical reference)	Documents, photos and photo albums, music, video *memory-aid, entertainment, medical history, progeny*	Books, papers, reference documents *memory-aid and reference*
Working (daily use)	Documents, email, photos, audio including CDs, video *communication, entertainment, financial records*	Documents, email *content for professional use to communication*

- Recall a Chicago hotel stay over the last 10 years or a restaurant or wine from a dinner in Paris about four years ago.
- Find a cancelled check or receipt.
- Show figures from papers on supercomputers during 1980–1990.
- Find articles and papers that mention Amdahl's laws, including the original articles.
- Recall email and letters to or about *x* about five years ago even though it was not specified to be a letter or email correspondence. List letters, recommendations, and papers written in 1989.
- Display an album from a fishing trip or taken during July 1999 on the TV set, or display all the photos randomly on a large flat-panel display.
- Play a set of selections on a particular computer or the home stereo.

CyberAll Storage

CyberAll is currently held in the Windows file system. A decision was made to not use a database. This was based on: variation of document types; cost to create and maintain database columns, keywords, or metadata; inflexibility of moving or modifying files in an established database; concern that any database is not a "golden" data type and hence is likely to become obsolete; a belief that programs should be able to automatically extract any relevant metadata (letters, forms); and the ability of ordinary indexing and searching to solve most personal needs. Items are stored in a relatively flat two- or three-level folder hierarchy with a few dozen folders in the first level and an average of four folders in the second level. A plethora of specialized music database programs manage the encoding, organization, and playing of CDs, music files and music sources. Photos represent a challenge. The photo collection is called the shoebox, and indeed has that flavor. My database colleagues have yet to convince me that they can do better than "grep" searching the free text or viewing thumbnails.

The author has also used descriptive file names to aid retrieval. A name might include subject, organization, keywords and a date. Many file types, including Word files and JPEG images have extensive metadata. Photos in JPEG format include title, subject, location, description, category, keywords, dates (taken, modified, and so forth), and camera information.

Documents are retrieved by searching file contents. For example, searching is instantaneous using AltaVista or the Windows 2000 file system. Eventually all the information of or about a file inherent in the file is needed. Systems need to "understand" the documents, for example, the letters and receipts they hold.

Photos

Photos are stored as individual photos in a set of personal and professional folders, and albums—when there is a story. Retrieval is by date, photo name or any other text attributes when they have been so labeled. Most of us are unwilling to label and describe each photo since a year of photos by a prolific amateur could take several days to label. Thus, the alternative is viewing a folder of thumbnails and using emerging image searching programs.

A photo (or a pointer—shortcut—to it) is stored in every folder where a user might expect to find it. Folders provide an organized, yet open-ended filing structure. Folders are grouped as: time-based events (trip, party, conference); and subjects (family member, hobby, mountain scene, food). One can easily have three attributes or folder sets where a single photo (or pointer) is stored, for example, French 1997 trip, French mountain scenes, and all mountain scenes that include France. Sunsets might get a fourth filing. Each time a new, useful category is found, thumbnails are made and inserted in an appropriate folder. Tools for compound searches such as mountain and sunsets would be useful.

Obviously, there are a plethora of functions that can be invented to facilitate filing, labeling, and retrieval. Speech input, for example, offers great potential to assist filing. Arcsoft's Photobase (see www.arcsoft.com/) creates albums with searchable keywords and audio segments for each image.

Capturing and Encoding Everything (Items and Formats)

Legacy data types such as CD, paper, photo, and videotape have stood the test of time and various tools have allowed them to be cyberized. In contrast, for computer-created items, the application program that created an item may often no longer be available, so items are essentially lost. Over the long term, older versions of complex programs like databases, word processors, and computer games may no longer run

Figure 1. The TIF format is the basis for most OCR and page-input programs.

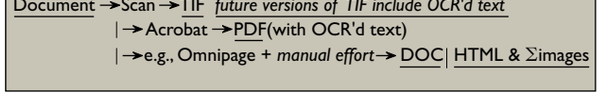


Figure 2. Example PowerPoint file conversion.

on new systems. Information must be held in a few golden primitive formats because these have to be supported forever—to date, only TXT format seems to be readable over decades. CyberAll documents are stored in at least two formats to increase the likelihood of reading the document in the future. Black-and-white documents are scanned and retained as TIF files and also converted to some OCR'd form, for example, PDF for retrieval. Some documents are converted to Word or HTML for searching, viewing, printing, and even editing. For example, a scanned copy of the 1889, 13-page Hollerith patent TIF file requires 700KB and 79MB for black-and-white and color, respectively. A PDF file of the image for limited on-screen viewing, printing, and searching is about 1MB. DjVu-stored (see www.DjVu.com) color documents appear to encode compound color and text documents in half the size of other formats. File formats such as JPEG, HTML, PDF, RTF/DOC and TIF are "golden" formats; PowerPoint is a container for photos.

Capturing paper documents. An HP Digital Sender was used to scan to either black-and-white or color TIF or PDF.[4] For most working documents PaperPort is used to scan to a TIF dialect with implicit OCR. It is difficult, though necessary, to cut a relatively rare bound book, paper, or report apart to scan and discard. Some documents (engineering notebooks and notes, for example) have not been scanned due to readability and contrast. If a document needs to be permanently preserved, it is converted to a golden format to increase its likelihood of permanency. The TIF format is the basis for virtually all OCR and page-input programs, as shown in Figure 1.

Various OCR programs can recognize and convert a document into a repurposed, near likeness of the original or even an HTML page. The MHT format, derived from MIME, can hold the collection of files for an HTML page in a single file. The evolution of TIF and HTML-XML to hold different image encodings, including recognized text, JPEG, and GIF objects will make scanning more convenient, economical and useful.

Future TIF standards include the image, OCR'd text for searching, and metadata (the various dates, author, and keywords that further describe the document). Scanners that directly connect to a personal computer usually just provide bitmap images and, depending on the interface software, images can be stored in a various formats.

Capturing photos and creating albums. Photos are scanned into folders. Albums hold stories such as a trip, birthday party, or a period of a family's life. PowerPoint is the main container for albums, but in addition the photos are retained in folders since one photo may appear in several albums.

PowerPoint can be converted directly to HTML format for Web hosting, or alternatively an HTML document containing the photos can be created using various Web-authoring tools. PDF albums are used to encode legacy paper albums (multiple pictures mounted on a page). All photos are JPEG; TIF is not used as the intermediate images format because of size. Kodak's photo CD conversion service, and Nikon and HP scanners were used for photo input.

Time and/or costs to scan and encode paper, photos, and CDs. As a rule, simple items such as a page, a photo, or a slide cost about one dollar from commercial services. Articles approximately 10 pages in length can be scanned directly into PDF in about two minutes with the HP Sender and captured in Acrobat format at three pages per minute using a 400MHz PC. Photo scanners require approximately 20 seconds to two minutes per photo. One may want to recognize a document and convert it to a perfect, editable document such as DOC/RTF or HTML. This requires "perfect" recognition together with the need to format the document exactly like the original. Such a document is being republished. To scan, recognize, and edit a page can easily require 10 minutes to create a formatted document that is suitable for repurposed use. The time to encode or "rip" a CD depends on the CD reader speed, tools, and availability of databases that can be used to create labels. CDs took roughly 10 minutes of attention time to read and label the tracks.

How Long Will a Data Format Remain Valid? (Consider 8-track Tape)

The most serious impediment to a lasting archive is the evolution of media, platforms, formats, and the applications that create them. Unique, proprietary, and constantly evolving data formats such as Acrobat-4, MPEG-4, Oracle 8, Quicken 2001, Real G2, and Word 2000 suggest or even guarantee obsolescence.

[4]The PDF format is a current, significant de facto standard claiming approximately one billion documents, implying a total capacity of at least 100TB.

Table 2. Storage requirements and cost for common data items.

Items	Size (Bytes)	Encoded size	Items/ GByte	Cost($)/ Item*
page (b/w) fax	100K	4K	10–250K	0.00004–0.01
page (color)	6M	0.3(JPEG)	160–3,500	0.003–0.06
business card	5K	500	200K	0.00005
photograph	3M	25–400K	10,000	0.001
book 350pp	25M	1–2M	40–750	0.01–0.25
CD (1hr)	640M	60M	1.5–16	$0.60
LowQ video/hr	50–300Kbps	20–300M	3.3–50	0.002–3.30
MPEG video/hr	1.5Mbps	670M	1.5	6.70
HiQ video/hr	DVD 4Mbps	1.8G	0.6	18

*2000 system prices of $10,000 per terabyte or $10 per gigabyte.

Table 3. Size for storing everything read/written, heard/spoken, photographed and seen (via video).

Data types	Rate (Bytes/hour)	Per day/ per 3 year	Lifetime amount
read text, few pictures	200K	2–10M/G	60–300G
email, papers, written text		0.5M/G	15G
photos w/voice @100KB	200K	2M/G	60G
photos @200KB	10 photos/day	2M/2G	150G
spoken text @120wpm	43K	0.5M/G	15G
spoken text @8Kbps	3.6M	40M/40G	1.2T
video-lite 50Kbps POTS	22M	0.25G/T	25T
video 200Kbps *VHS-LITE*	90M	1G/T	100T
DVD video 4.3Mbps	1.8G	20G/T	1P

The new version may not read legacy data on legacy platforms forever. The basic question is: "How will the data be readable in 10 or 50 years—what are the few, 'golden' data formats that we can depend on forever?"

Since CyberAll will store all personal information, including documents, photos, and videos, this data needs to be valid and hence understood in an indeterminate future! High-quality paper will hold information for a millennium (or at least several centuries), and film is sometimes rated at several hundred years (if you keep it very cold). A CD is likely to be readable in 50 years, but finding the CD reader/computer and file system/app to read it will clearly be impossible if history is a guide.[5] Is paper the only true long-term storage medium?

Digital documents are committed to a conversion treadmill. With each generation of media (circa 1978 8-inch floppies), the computer system (CPM), and the application (Wordstar), a conversion is required. This happens about once a decade, if we pick formats carefully. For plain documents, the alternative is paper stacks of personal information in file cabinets, as compulsive information pack rats do today, versus a single DVD that a computer can search. The JPEG format is constrained by camera equipment and the need to interoperate; TIF is constrained as a facsimile standard. Starkweather's Pedistil system [7] scans documents (books, journals, papers) into TIF at 400dpi followed by OCR for retrieval.

For data to be understood in the future, it cannot be subject to applications that change every year such that a particular version has to be maintained (for example, Quicken 95–2000).[6] As applications evolve, this means data maintains the creating version of the application or all past data associated with a named application has to be converted forward. This is also an issue and perhaps failure of object technology that runs on a single, universal machine.

Alternatively, the one way to ensure interpretability is to transform data emanating from a program, into a generic format. The current solution for longevity is to use a few widely accepted data types that data is transformed into. For example, yearly Quicken reports end up as text files.

Given the vast amount of data in Adobe's PDF, or Microsoft's Office[7] what commitment will the apps make to their data? Will the JPEG and TIF working group ensure that my old files can be read?

Economics

Table 2 gives the storage requirements and costs for various data. The cost of documents and photographs in a CyberAll is nearly zero, hence purging anything is generally a bad strategy. The cost for storing encoded CDs is about 1/20th the cost of a CD, not counting the encoding time. The encoding cost is comparable to a CD cost. Playback appliances and personal computers are likely to change the music distribution industry, making CDs obsolete within a decade.

Table 3 estimates the requirements for storing an individual's content for life. One will be able to record all of the information accumulated in their entire personal and professional life in a few terabytes, including everything spoken, but not everything captured via video. This archive would include all home videos for most families. The table shows the various jumps in storage required going from recording lifetime text, transcribed or encoded speech, and video. The need to

[5]3000 of the author's documents circa 1975 stored on 8-inch floppy disks created on a Digital PDP-8 word processing system were converted into Microsoft Word format using a PDP-8 emulator running WPS 8 software.

[6]Data written in 1990 on a Macintosh and converted forward to a more recent version is unable to generate a report. Data written on a Macintosh cannot be converted across (that is, read) on a PC without manual effort. MacDraw and Draw for the PC have similar problems.

[7]Data written in the early 1980s can been converted forward from Macintosh versions 4.0 and converted across to the PC Office 2000 standards for Excel, PowerPoint, and Word. PowerPoint files can be converted to JPEG or HTML formats.

Table 4. Comparison of the author's document, photograph, videotape, and 150-CD archives.

What	Files	Size(MB)	MB/file	MB/Yr
Digital document archive	8234	1629	0.20	50
TIF and PDF paper scan archive	3402	2788	0.82	1000
GB books (4 encoded)	2027	494	0.24	
Photos (all types)	3920	492	0.13	50
Photo albums (PowerPoint)	55	151	2.75	
Mail (2.5 years)	2	330	165	200
GB videos (lectures, 8mm)	20	4000	200	
150 CDs MS multimedia @16Kbps	1497	5820	3.89	600
	19157	15704		1900

recognize and only handle transcribed speech is clear based on storage and searching needs.

The actual storage (Table 4) of my CyberAll is quite small. Until recently I purged paper files to stay within cabinet and searching constraints, hence my CyberAll is incomplete. These are 2–15MB PowerPoint albums of JPEG photos.

Deciding among the array of programs using mostly proprietary data formats and developing a process to deal with the encoding of documents for personal/professional and archive/working use is probably the most difficult decision in building one's CyberAll. These formats have to be maintained or converted in the future. One strategy is to wait for an ideal solution and stability that will hopefully come within five years. Alternatively, keeping data in the most primitive, scanned or encoded form—TIF and JPEG—allows for future flexibility, including being able to utilize better tools, such as OCR and encoding.

Photo storage and retrieval are certain to improve. Cameras and acquisition software must maintain dates. Cameras should include audio recording for voice annotation, metadata for retrieval, and improving the value of the image with a bit of sound or voice. Advances in query software to find like images such as buildings, people, or sunsets using color spectra, shapes, and other attributes is improving to the point of usefulness. It is necessary to widen CyberAll's scope to include all family members in order to get a better handle on everyday archiving and use for noncomputer users. My CyberAll operates with and is aided by my file organization. A general system to operate over decades would have to posit a structure and tools to aid users and cope with the intergeneration problem. CyberAll requires tools that relate to privacy, especially the ability to lock files until time has elapsed or events occur. And the ability to limit visibility of documents to specific people or people in specific roles.

In the future, retaining conversations and video are possible. Devices to record interviews and meetings would be welcome and necessary. When conversations are added, ownership and privacy become even more complex. Video will be the focus for many users' CyberAll. Waiting for better editing, searching (especially using recognized text), and retrieval tools and even larger disks is my strategy. Already, the project has convinced me that a goal of paperless storage and transmission is attainable now for everything except books and items that represent money.

Conclusion

In 2000 the cost to store all personal and professional related, computer generated and paper forms of information, including CDs and photographs is nil, especially compared with physical objects. It is costly to load a personal store and to maintain it for the indefinite future. Information is held in multiple formats to increase the likelihood for long-term retrievability. Scanners to TIF with text OCR will make paper input as easy as discarding it. Thus, a state of paperless storage and transmission is near. Standards and ease of use are now the key enabling technologies.

In the next five years, anyone will be able to have a personal computer that retains everything they've read, written, and presented via video that originated from a computer or legacy source such as paper or videotape. This would include all of the transactions for a family, ranging from general correspondence to every conceivable medical and financial record. In 10 years, systems should be able to recall every personal lifetime conversation. Currently, significant effort is required to build and utilize such systems if they involve the entry of legacy documents—consider modifiying and organizing collections of books, papers, photos, or videotapes. A system such as CyberAll could quite possibly be a killer app for personal computers. C

References

1. Bell, G. Dear Appy, How committed are you? Signed lost and forgotten data. *ACM Ubiquity* (Feb. 2000); www.acm.org/ubiquity/views/g_bell_1.html
2. Bush, V. As we may think. *Atlantic Monthly,* (July 1945); www.isg.sfu.ca/~duchier/misc/vbush/vbush.shtml
3. ECabinet Product Brochure. (Nov. 1999), Ricoh Corporation; www.ricoh-usa.com.
4. Gates, B. *The Road Ahead.* Penguin Books, 1996.
5. Huhns, M.N. and Stephens, L.N. Personal ontologies. *IEEE Internet Computing 3*, 5 (Sept. /Oct. 1999), 85–89.
6. Lesk, M. *Practical Digital Libraries.* Morgan Kaufmann Publishers, San Francisco, 1997.
7. Starkweather, G.K. *Pedistal: A Personal Document Imaging System.* Microsoft Research Technical Report MSR-TR-2000-103, June 2000.

Gordon Bell (gbell@microsoft.com) is Senior Researcher at the Microsoft Bay Area Research Center in San Francisco.