the future:

digital documents

chapter three

I must confess it feels funny to be writing about the future of digital documents... in a book! And not a dynabook, either.

I believe that books will outlive their original "purpose" in any case, just as we still keep our dogs though we may not need them for hunting anymore. I think books will be around for a long time.

But the dynabook's time is nearly upon us, and information can be better accessed in digital format. Even in this seemingly ephemeral format of a book, if we can not hope to view the future itself, at least we point to the places to look for the future. Like stars on the horizon, there are reliable pointers that will last through the years. This chapter points to those constellations most likely to continue to appear in predictable revolutions.

The Fantastic Pace Of The Web

Ray Kurzweil is the modern-day Thomas Edison. Ray has invented machines that can read, that can understand speech, and that have the ability to learn! The philosophies behind his inventions of 20 years ago are applicable to digital information today in our quest for instant access.

> *"Moore's law states that computing speeds and densities double every 18 months. In other words, every 18 months we can buy a computer that is twice as fast and has twice as much memory for the same cost.*
>
> *"Moore's law actually is corollary of a broader law I like to call Kurzweil's law on the exponentially quickening pace of technology that goes back to the dawn of human history. I mean not much happened in, say, the tenth century, technologically speaking. In the eighteenth century, quite a bit happened. Now we have major paradigm shifts in a few years' time.*
>
> *"Computer memory is 150 million times more powerful for the same unit cost than it was in 1948, the year I was born. If the automobile industry had made as much progress in the past forty-five years, a car today would cost about a hundredth of a cent, and would go faster than the speed of light.*
>
> *"Moore's law is providing us the infrastructure in terms of memory, computation and communication to embody all of our knowledge and methodologies and to harness them on inexpensive platforms.*

"It enables us to live in a world today in which all of our knowledge, all of our creations, all of insights, all of our ideas, our cultural expressions, pictures, movies, art, sound, music, books and the secret of life itself are all being digitized, captured and understood in sequences of ones and zeroes.

"Thus around the end of this decade, a full print-to-speech reading machine will fit in your pocket.

"And Moore's law projects that our personal neural computers will match both the memory and the computational ability of the human brain - 20 million billion calculations per second - by around the year 2020.

"In the year 2040 ... In my view, Moore's law will still be going strong. Computer circuits will now be grown like crystals, with computing taking place at the molecular level.

"By the year 2040, in accordance with Moore's law, your state-of-the-art personal computer will be able to simulate a society of 10,000 human brains, each of which would be operating at a speed 10,000 times faster than a human brain.

"Or, alternatively, it could implement a single mind with 10,000 times the memory capacity of a human brain and 100 million times the speed.

"What will the implications be of this development?"

Excerpts from "The End of Handicaps," Ray Kurzweil's Keynote Address at the 1996 International Conference of the Association for Education and Rehabilitation of the Blind and Visually Impaired

Ray Kurzweil's work can be found on the Web at

http://www.kurzweiltech.com

Artificial Intelligence offers the greatest promise for automated recognition of information in unpredictable data streams. The combination of OCR and AI was coined as ICR. AI is also implicit in Intelligent Agents.

Surfing In Waves Of Information

In later chapters we discuss how to narrow down Web searches to achieve the most efficient information retrieval. We have a constant need to find better ways to navigate the ever larger ocean of information on the Web. Even in narrow fields, the Web spawns new information sources every day, and it becomes a daily, time-demanding effort to review all of the latest postings and publications.

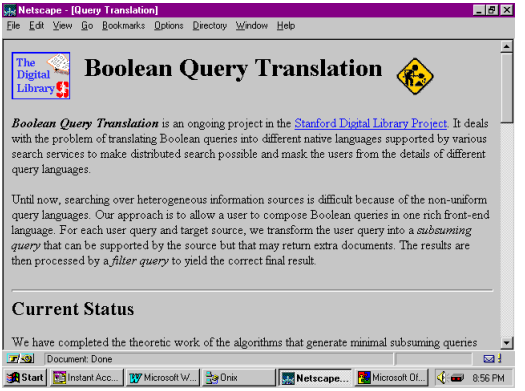
tip

On the Web, geniuses and charlatans have equal access. As always, the audience and the market at large are subject to feedback. Valuable sites offer the user real content, and hype sites won't be visited twice. Intelligent agents can make the first visit and save their "masters" the time.

Before the Web search engines were toddlers, the overwhelming volume of info was recognized as a major user stumbling block. The very search engines that were taking users into this cataract of information quickly transformed themselves into brand-new vessels that could navigate the torrent.

The Web search engines were the first rafts or rowboats that allowed users to go out upon the raging waves of information on the Web and actually make some headway. When the population and density of Web sites grew explosively, the need arose to automate the search functions that people need to perform repeatedly. Intelligent agent software is designed to provide a precisely focused info-gathering robot for every user, providing searching automation to save the user time and allow people to concentrate on ideas rather than on running programs.

http://www.diglib.standord.edu/dilib/



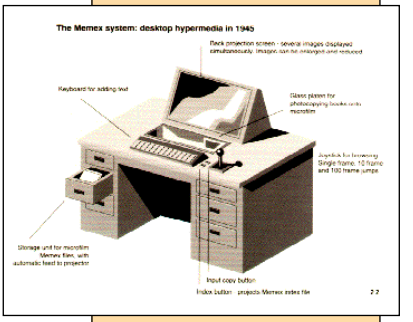
Herculean efforts are being devoted to make it easier to find information and make it less like using a computer. The goal of the Boolean Query Translation project is to create one rich language that can operate on all search engines, freeing the user from the tasks of learning many Boolean Query styles.

Intelligent Agents

The 007-ish name intelligent agent is a partial solution for the need to wade through piles of digital information. The average Web user can't and won't put in a lot of time to take advantage of the world of information on the Web. It is already increasingly difficult to keep up with the information streams pouring into the Web, and it's growing by magnitudes each year.

Intelligent agent software automates your surfing in these floods of future information, and many personal information robots are now available on the Web. Some of them can even "watch" what you retrieve and automatically retrieve "more like that." Be careful whenever this feature is available because improper use can waste rather than save time.

THE IDEAS ARE NOT NEW: THE MEMEX IS VISION OF WEB



Memex GIF of early illustration of imagined Info Retrieval Machine. Note the early mention "web" in Vannevar Bush's thinking.

This segment from Bush's original article presciently described the Web today:

"The human mind...by the association of thoughts, in accordance with some intricate web of trails carried by the cells of the brain... trails that are not frequently followed are prone to fade...Yet the speed of action, the intricacy of trails, the detail of mental pictures, is awe-inspiring beyond all else in nature."

Vannevar Bush, describing his fantastic Information Retrieval machine, which he dubbed the Memex in "As We May Think" in THE ATLANTIC MONTHLY, July 1945.

Never surrender categorization or gut judgment to the software; always employ common sense. Agents must be constantly tinkered with to achieve best results.

Then consider the matter of bandwidth. Those old copper wires that have been hanging from poles since the days of the telegraph still hold a lot of potential. Using currently installed copper wire, ISDN reliably delivers point-to-point communications at four times the speed of the latest conventional modems. ISDN has been available for years, and ISDN modems are relatively affordable and practical.

Through 1996, some telecomm service providers charged time-based rates. Even two cents per minute adds up when the remote work station is constantly connected, as they will be in the future, and as some are now. This ISDN service is controlled by the telco, and other than that, no network changes are necessary, except for the special modems on each end. It should be no surprise that the AT&T spin-off called Lucent makes a modem that can run data over the same wires.

Service providers are being called a lot of things lately, so here's some clarification on the matter:

Telco usually refers to the phone companies, from the Bell Atlantics on down to the local providers who overcharge you at pay phones on the highway.

Telecommunications service provider is basically the same thing but refers to digital rather than voice communication, and includes everything from your local phone line to the long haul lines and backbones that make up the Internet, including Sprint, MCI, et. al.

An Internet service provider (ISP) maintains a constant link to the Internet by whatever means, from 56 KB SprintLink to leased T1, T3 lines and so on. The key is that an ISP provides a constant link to the Net, and many users can use the ISP to get to the Net.

WHY LOOK TO THE PAST IN A CHAPTER ON THE FUTURE?

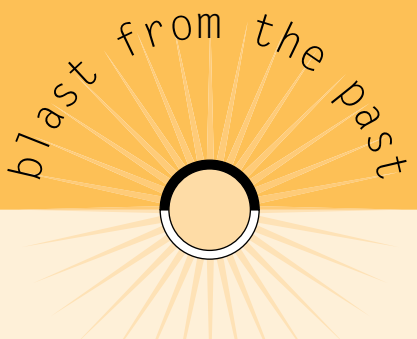
To give you a sense that you can see the future; it's not a hopeless blur. Vannevar Bush and Bucky Fuller and John Warnock and Ray Kurzweil and Kelly Johnson and Leonardo da Vinci—they all saw it. Clear as a bell. For example, look at the quote from Bush on the previous page describing a browser. Any Web surfer can choose how long his previous Web links stay highlighted by setting preferences in his browser; it's equivalent to Bush's quote of "trails that are not frequently followed are prone to fade." And, Bush's idea of "next that is suggested by the association of thoughts" is exactly the way concept search engines like Excite do their thing. It's not a stretch—it's an exact prediction of the Web.

And Ray Kurzweil's earlier comments confirm Bush's idea from the past to be currently realized in the present, and expected in the future.

Bill Gates, the leader of Microsoft, addressed the question of bandwidth in a very forward-looking way at Sapphire '96, the SAP User Conference in Philadelphia. Gates confidently reassured the audience that we won't run out of Internet, we can depend on it growing with demand.

"Four years ago, 80 percent of desktop applications were stand-alone, such as word processors, spreadsheets, databases and so on. Today, 80 percent of desktop applications are Microsoft Office or Professional, " Gates said.

The implications for the server software market are very clear. As Dr. Hasso Plattner, Co-Founder and Vice-Chairman of SAP pointed out in his discussion of current trends, "The majority of our applications are still running on UNIX, but NT and the AS/400 are now taking 50 percent of our new installs."



Before the anti-trust breakup, a nationwide network of connections seemed like a good idea. AT&T, which might be considered an artifact of early 20th century thinking that it was, really worked. Really old folks remember when there was one telephone company, The Bell Telephone company.

Megalithic thinking has its virtues, especially viewed in light of the pragmatic triumph of the Internet.

Mr. Gates stated, “The world of personal productivity explains the popularity of PCs in business. In the past, information was just printed out, and there was no strategic use of the information. MS Office provides tools for evaluating information and doing What Ifs.” All of that required more and more bandwidth.

Bill Gates predicts future innovations, stating “no time frame, but it will happen”:

- Handwriting and video input
- Computers will talk, see, listen and learn
- 3D will become commonplace for collaboration
- No doubt the Internet will be the primary communications tool

You Are Not A Lonesome Pioneer

If you are going to check out only one of these global taps, try the Digital Libraries Research and Development page. Don’t be shy; this page is dedicated to equal rights to information. Feel at home.

This page tracks many digital library projects, hence the “dlib” in the URL

http://www.dlib.org/reference.html

Digital Information In Perpetual Action

The best way to look toward the future is to experience it yourself, firsthand, via the Web or an Intranet or someone else's network. The following Web sites provide a fascinating glimpse of technological innovation applied to information accessibility.

The Gutenberg Project

http://promo.net/pg/history.html

> *"The Project Gutenberg Philosophy is to make information, books and other materials available to the general public in forms a vast majority of the computers, programs and people can easily read, use, quote and search.*
>
> *"Alice in Wonderland, the Bible, Shakespeare, the Koran and many others will be with us as long as civilization ... an operating system, a program, a markup system ... will not."*
>
> *Project Gutenberg Web site*

The title of *www.etext.org* is Electronic Books, and this site is dedicated to a very special mission. When books go out of copyright time obligations, this group transfigures books into universal text that will live on in the digital future. To do this, they reduce everything to plain ASCII.

http://www.promo.net/

(which plays soothing background music while you are on the site)

http://www.etext.org/books.html

http://www.w3.org/pub/WWW/Protocols/

(for background on how it all started)

tip

Which three languages will be on the Rosetta stone of digital documents?

What would work at the moment? Which are the most popular and robust languages to give generations hundreds or thousands of years in the future the chance to decipher our many formats?

Right now, it's ASCII (including HTML and SGML), Microsoft RTF and Adobe PDF.

(With all due respect to the only language explicitly designed for this purpose of long-life archives, SGML is ASCII in a precise syntax.)

The Digital Library Project

"At the heart of the project is…a uniform way to access a variety of services and information sources."

From http://diglib.stanford.edu/

This arm of the Digital Library Project is linked to all of the other nationally connected branches of this future-directed initiative to move today's information onto today's and tomorrow's media. Ongoing, current information can be always found through these links. This nationally enriching research project is funded by the National Science Foundation, the source of many of the great ideas and technologies of the last 50 years.

HOW IS ALL THIS CHANGING PUBLISHING AS WE KNOW IT?

The development of extremely sophisticated electronic presses, namely today's common laser printers, allows an epochal paradigm shift. We have gone from print and distribute, to distribute and print.

Rather than fund and maintain a centralized printing facility which creates paper output, digital distribution of printable files for local paper output offers excellent cost benefits. Rather than print 10,000 manuals, give users the ability to print manuals with their own laser printers. Perhaps only 500 manuals are ever printed, because most users rely on online user guides and individually print out a few pages as needed.

When an idea such as this is embraced and espoused by everyone from Adobe to Xerox, it takes on the reliability of common sense.

The United States funds and conducts the lion's share of research and science in many fields, from aerospace to superconductors, and NSF is the government's proud masthead for these successful adventures.

http://www.dlib.org/reference.html

This site maintains clearinghouses for digital library research:

United States National Information Infrastructure Virtual Library, Library of Congress, World Wide Web Virtual Library, HyperDOC: A Service of the U.S. National Library of Medicine, and MedWeb.

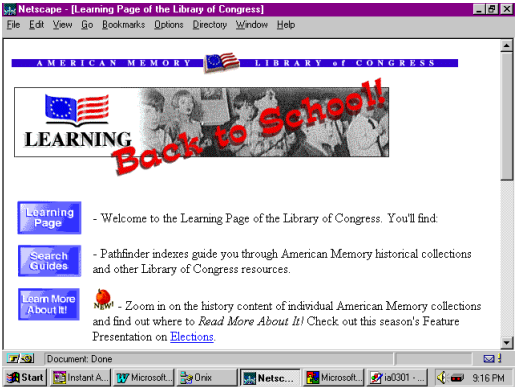
For the latest research, the ACM Special Interest Group on Information Retrieval (SIGIR) maintains pointers to digital library research projects, technical papers, conference announcements and proceedings, and calendars of events.

The Electronic Journals Publishing site maintains lists of electronic journals, related projects, papers and discussion lists.

Library Of Congress

http://lcweb.loc.gov/

The National Digital Library Project is truly a noble endeavor of our government, proving that it can do some things right. This project is specifically designed to fulfill the lofty aspirations of Bucky Fuller's Education Automation Dreams. This is a very uplifting use of technology for the common good. It helps people learn by providing not only access, but also the tools to catch the fish of information. It is the idea that it's better to teach a person to fish than it is to offer a simple meal. One is transient, the other is forever nourishing.



Computers are not intuitive to all of us, especially those of us born before 1980. This page leads to kindergarten through graduate school education on information retrieval on the expanding body of knowledge. To experience this first hand, visit:

http://lcweb2.loc.gov/ammem/ndlpedu/

Center For Electronic Text In Law

http://www.law.uc.edu/CETL/

At the Marx Law Library of the University of Cincinnati, an ambitious project is underway to digitize a collection of unique documents, to provide both electronic access and electronic preservation. Project Diana is named in memory of Diana Vincent-Daviss, a pioneer in these endeavors at the Yale University Law Library. Under the direction of Nick Finke, J.D., a carefully selected array of tools performs the tasks of scanning, OCR, SGML encoding and publishing on the World Wide Web.

Ignoring all of the technology for a moment, “you have to remember that we are a library,” Nick emphasizes. “Our goal is to get documents to people who have a hard time getting them, and to help them find critical information in poorly indexed documents.” In this effort to assist legal scholars, most of the issues of the digitization of paper information have been encountered and engaged successfully.

While the Web looks like an rich smorgasbord now, these early years will be like the first decades of flight. They will look like Orville and Wilbur Wright's first efforts. No matter how undeniably successful, this must be the most prehistoric phase. Our technology will soon catch up to our ideas, and then we'll have to improve our technology again.

For example, Kelly Johnson at the Skunk Works had to invent tools to work the new material of titanium to achieve the performance of hypersonic aerodynamic designs. Right now, new tools for machining the titanium of higher-speed access are being inexorably improved.

Three years ago, the forests of information on the Internet were comparatively impenetrable, and now there are super-highways through vast resources. Three years ago, indexes of files and directories were searchable through Gopher, Archie, Veronica and WAIS. These were demanding applications to learn and master. Today, even the greenest beginner can use Web search engines to comb through the very contents of the Web.

"We started out with the idea of preserving a precisely indexed collection of images," Nick explains, "but we came to the realization that we needed to create electronic books, not just images in a database." Working in conjunction with Don Waters and Project Open Book, it became obvious that it was critical to represent the intellectual structure of the book as well as scanned images of pages.

Anyone who has used computers for more than a couple of years has a sense for how rapidly generations of hardware and software come and go. On the other hand, documents maintain importance for years. In the case of libraries, the knowledge in books will be important forever.

"The organizing metaphor should be pages in books in a library, not pages in a folder on a desktop," states Nick Finke, in what should be a clarion call to others engaged in the task of document digitization. Producing "industry standard" .tif files is only a rough replacement for microfilm; building globally accessible libraries is another task with another set of goals. One is archival, the other is a form of re-publishing, or perhaps more accurately, re-broadcasting the data.

tip

To answer that always nagging question, "What's Legal?," this site offers an enthusiastically updated source for legal opinions, focused on intellectual property concerns. The Web and intellectual property seem to be at odds because one offers universal access and the other requires individual identity and copyright protection. The answers are hard to come by, but look here for starters.

What can I copy off the Web?

What can the Web copy off my site?

http://www.findlaw.com/01topics/23intellectprop/index.html

The law doesn't have to be a mystery when it comes to digital documents. This site offers current, continuing research and opinion on the subject of intellectual property in the new media.

The American Memory

The Library of Congress is very snazzy these days, and the American Memory Project is the embodiment of many predictions of what a World Wide Web could do. Visit the site at:

http://lcweb.loc.gov/

http://www.w3.org/pub/WWW/

The Web is truly global, with supporters on all continents.

Always a good spot to check out:

"The World Wide Web (W3) Consortium exists to realize the full potential of the Web." This consortium is funded by both industry and government sources and offers updates on the latest developments in all phases of the Web."

With a gift from Ameritech, the Library of Congress is sponsoring an open competition to enable public, research and academic libraries, museums, historical societies and archival institutions (except federal institutions) to create digital collections of primary resource material for distribution on the Internet in a manner that will augment the collections of the National Digital Library Program at the Library of Congress. The National Digital Library is conceived as a distributed collection of converted library materials and digital originals to which many American institutions will contribute.

The Library of Congress' contribution to this World Wide Web-based virtual library is called American Memory and is created by the National Digital Library Program.

In the 1996-97 competition, applications were limited to collections of textual and graphic materials that illuminate the period 1850-1920 and that complement and enhance the American Memory collections already mounted in the National Digital Library.

Making Digital Documents Better Than Paper

Just as PostScript was the cornerstone of electronic publishing, enabling desktop tools to create superb creative pages, Acrobat serves this function for the documents of the future that will travel over the global Internet. Adobe PDF format offers fidelity across paper, the Web, CD and multimedia.

The Portable Document Format is much more capable of going out on its own in the broad world of unpredictable platforms than is PostScript. PostScript was a programming language that precisely defined the layout of ink on a page and had to be executed by an interpreter on the printer side.

As PDF becomes a standard page-definition language, all new printers will be able to support the rich output format (within the limits of the hardware) because no interpreter will be required on the output side.

The PDF Group is a consortium representing commercial printers who currently form a $10-billion-per-year industry. The PDF Group was formed to advise Adobe on future development of the Portable Document Format. To provide solutions for the extremely demanding requirements of this industry, which include both highest quality and highest volume output, the Acrobat family will continue to evolve.

In addition to Acrobat, Adobe Systems' other products are all evolving into both paper

and digital document-creation tools. “Repurposing” is the name given to the process in which formerly print-only documents and processes advance to provide information on multiple media. PageMaker, for example, can be used to create documents intended for paper and digital form, and the digital form may be either HTML or PDF. Similar evolution is far along with Illustrator, FrameMaker and other Adobe tools.

But the real key is access. Already the Lycos and Yahoo pages reference many other full text search engines, and offer push-button access so a user can execute a query on one or more search engines.

In the commercial world of Information Retrieval, vendors such as Excalibur, Fulcrum, Open Text, Verity and many others offer the ability to execute a query on many servers and indexes simultaneously. On a Wide Area Network, a single query could interrogate all online corporate information assets. And to accommodate the user, the software blends all hits from all servers into one relevancy ranked list.

Right now, everything from handmade Java scripts to high-priced services offer the ability to run searches across multiple search engines.

The best site for managing these tools at the moment is the User. Right now the dynamic feedback and discrimination of an every day online user is superior in many ways to info robots.

To serve the user, the search engines will increasingly offer easily understood and easily reconfigurable hit lists. It is these hit lists, properly designed to convey the most important information, that will enable individuals to pick intelligent paths.

When you have become familiar with the results of searching, you get comfortable with setting up your persistent searcher, your intelligent agent.

We are on a precipice of technological development here. Artificial intelligence is not a proven commodity, not at all. We are at the point in a dramatic technology where aerospace arrived in the 1960s. In the twenty year interval between the successful employment of jets and the setting of the unbreakable speed and altitude records of the SR-71, vast advances were made. In the thirty years since, no aircraft has come close to matching that early burst of technological genius, embodied in the Mach 3 Blackbird.

We may be at a similar point in Information Retrieval, or we may not.

In any case, the vast information access now available is like the cheap air travel that arose out of all this jet testing. So we only fly 500 miles per hour from city to city now and not Mach 5, is that so bad? Compared to driving at 55 or 65 mph, air travel is in another dimension. You could never consider a one day overland trip from Philadelphia to Chicago and back. Via the airlines, it's a routine hop.

TWICE THE SPEED FOR DIAL-UP SERVICE

Lucent Technologies recently announced new chip technology that will allow PC users to access the Internet almost twice as fast as ever before. At 56 KB per second, you'll see dramatically improved use of Internet applications, especially downloading graphics, video conferencing and collaborative computing.

For more information about Lucent's technology, go to:

http://www.lucent.com/Whatsnew/whatsnew.html

Today's Web information retrieval engines offer infinitely greater convenience to learning than even jets offer over driving. You, your kids, your parents and people all over the world have access to this global library of knowledge.

Now that information has been recognized as the world's most valuable resource, we as users can expect ever better and cheaper access to information. The global spread of television is an early technological testament to human curiosity and desire for knowledge. The dawn of a library at every table will be a nurturing boost for mankind, allowing every individual to pursue their own interests and creativity.

Summary

As explained by Ray Kurzweil and Moore's Law, computers double in capability every 18 months, and this trend will continue. The computers themselves will continue to become faster, offer greater capacity and user conveniences, and either stay at the same price or drop in price.

Access to high-speed communications is likely to be as hotly contested in the future as long-distance services have been in the past. The consumer will benefit with faster connections and cheaper prices.

A few companies were poised in certain dominant positions at the start of the new media revolution, and they early on committed resources and strategic planning to the emerging trends. For example, Adobe and Microsoft have leveraged their original technology to be most effective under both current and future systems.

There are a number of projects underway on the new global information network that will track the progress of the latest developments as they are adopted by corporations, government, universities and libraries. Anyone interested in the state-of-the-art of these evolving systems and technologies can quickly update their understanding via the World Wide Web.

part 2

managing digital content