The Digital Information Revolution:
The Era of Immediacy

Michael B. Spring

May 2011

Table of Contents

Chapter I: Introduction

Origins of This Book

In 1991, I wrote a book entitled "Electronic Printing and Publishing: The Document Processing Revolution." I was driven to write the book by my excitement with the new systems which were emerging to process information, particularly large documents. There were several things in that first book that bore the test of time, but there are many more things that turned out differently. Most notable was my inability to anticipate the impact of the World Wide Web. While I anticipated hypertext systems and collaboration over the network, I missed the magnitude of the impact of the World Wide Web.

In making the decision to write this book, I thought long and hard about whether I was going to be able to say anything with more certainty than I had in the first book. The last thing that I wanted to do was add to the glut of information-limited publications – one of the not so desirable side effects of the World Wide Web. Some of the points I made in 1991 still ring true, and some of the thinking processes that led to those conclusions have solidified in the last two decades. These small increments in my personal knowledge help to clarify what is sometimes a chaotic set of advancements. Perhaps most importantly, students continue to come back and tell me that what I addressed in my lectures had a significant impact on how they viewed what is going on and continued to make a difference in how they approached work. It is reflection on the ideas that have stood the test of time that have motivated this book.

The leitmotif in this book is the impact of digital information. There are aspects of digital information that seem to suggest it is not the same as normal information. The book starts off looking at the history of how we communicate and exchange information. We have been exchanging information in symbolic form for thousands of years. About 150 years ago, we began to exchange information electronically and in some cases primitive digital form. Today, information in digital form is the norm. In its current digital form, the computer is currently working with me to correct grammar and spelling, to keep cross references accurate, and in the future, the system will work even more intimately with me to process the digital information. Another aspect of digital information relates to theft of information. People have always been involved in information theft, but that involved going though garbage to collect credit card receipts. Today, information about millions of people can be stolen in the blink of an eye, because the information is stored in digital form. The story in this book is the story of digital information and how it changing our lives. The most fundamental change is in how we communicate – suggested by the subtitle of the book "the era of immediacy."

Audience for this Book

This book was inspired by the many students who have told me they learned something important from my lectures. Students have driven me to stay abreast of technology and to keep learning. I find that students really want to understand technology at a fundamental level, but the way we teach at a graduate professional level often makes that knowledge less rather than more accessible. In addition, many of my brilliant colleagues fail to communicate to students the very intense excitement that drives their research. While I am getting too old to learn many more new things, I take great pleasure in trying to present difficult concepts in a way that makes them clear and that communicates the power of what can be done with the concepts. These lectures, which I refer to as "Mr. Wizard lectures",[1] have found their way into this book in various forms. This book is written for people who are curious about those aspects of our modern world that pertain to information. I hope, that like Bob Lucky's wonderful book **Silicon Dreams**, it is accessible to individuals who are not technically sophisticated and still of some interest to young and old professionals in the field.

The Author's Perspective

One of the best parts of writing a book is the opportunity to learn and reflect. A lot of the reflection has been looking back over more than thirty years of research, teaching, and writing. There were interesting resonances that were right on target, and some that were so far off as to be embarrassing. The core perspective in this book has been in my writing for all of those years and makes me more confident about the focus on the significance of digital information. The other part of the fun is the focus on

[1] The reference to Mr. Wizard is a tribute to Don Herbert who hosted a show "Meet Mr. Wizard" through the 1950's. It is interesting to reflect on the format of the show where Mr. Wizard invited young boys into his apartment to explain basic science in terms a child could understand. (It is a shame that we have lost that period of innocence.)

learning new things in the process. On the one hand, I was amazed at how many of the things that I suggested were around the corner five or ten years ago are already here. It reduced greatly the number of things that are in the book, because we have already gotten there. It challenged me to think about what would be next. As I looked at those predictions carefully, not wanting to find them in the market place before the book, I was amazed to find that some of them were already here.

At the core, this book is about a revolution in how we communicate – to suggest a phrase, "the era of immediacy." What I see as a "revolution" may be a mere hiccup in technological evolution, but I think not. If I am correct, the future will be radically different from what we know today. The revolution is not about printing and publishing. It is not about the World Wide Web. It is not about e-business. It has to do with the nature of human thinking and memory and our ability to communicate what we know and think to our contemporaries and to succeeding generations. In this sense, the revolution we are in the midst of today is similar to the revolutions caused by speech and writing. The proposition of my first book was that the revolution was similar to the one caused by the advent of mass production printing. The proposition of this book is much more audacious, i.e., I suggest that we are in the midst of a revolution in how we communicate.

Various authors from Freud to Adler to Maslow have discussed what it means to be human. There are several less well known views, such as those of Viktor Frankl who makes a compelling case for a "will to meaning." Johan Huizinga makes a compelling case in his book "Homo Ludens" for man the game player. One might make a case for story telling as a fundamental human motivation, and there is some scholarly work that draws a more complete exposition of that thesis.[2] There is no doubt that humans are driven to communicate and our competence in communication is what makes us different from other species.

The disruptive technology of our time, digital information processing via computers and networks provide us with an opportunity to make unprecedented improvements in how we communicate. In 1959, the French version of Teilhard de Chardin's **The Future of Man** was published. It consists of a compilation of pieces he wrote over the years about the development of global consciousness. Teilhard observed:

---

[2]Denning, Stephen (2000).The Springboard : How Storytelling Ignites Action in Knowledge-Era Organizations Butterworth-Heinemann.
Williams, Meridith (1989). Vygotsky's social theory of mind. Harvard Educational Review, 59(1), 108-126.

> *I am thinking of course, in the first place of the extraordinary network of radio and television communications which, perhaps anticipating the direct syntonization of brains through the mysterious power of telepathy, already link us all in a sort of "etherized" universal consciousness.*
>
> *But I am also thinking of the insidious growth of those astonishing electronic computers which, pulsating signals at the rate of hundreds of thousands a second, not only relieve our brains of tedious and exhausting work, but, because they enhance the essential (and too little noted) factor of "speed of thought," are also paving the way for a revolution in the sphere of research.*[3]

Teilhard's philosophy/theology puts great stock in the ability of humans to be conscious of each other and the world. Communication and awareness are what most distinguish us as humans, and we strive to become more aware and more interconnected with each other. If there is a leitmotif in the story of the evolution of computing, networking, and the web, it is that we will support and encourage the development of the technology in ways that better enable us to communicate and increase our awareness of our world and other humans.

Acknowledgments

There are many people to acknowledge and thank. But first and foremost, is thanks to the many students over the years that it has been my pleasure to teach and work with. I would also like to express my gratitude to my teachers. Fr. Aldo Tos introduced me to a variety of great authors while I was still in high school including Teilhard, Bonhoeffer, Camus, Sartre. Stanley Barondes introduced me to philosophy and humanism more generally while I was in college. Bradley Seager brought me to Pittsburgh to complete my PhD and taught me to write – or at least he tried. Over the years, I have been fortunate to have worked for people who were as much teachers as superiors. Sam Deep mentored me through my first fifteen years and helped me to understand how important it is to believe in the people with whom you work, and perhaps more importantly, how to view work as a collaborative effort. Over the next fifteen years, James Williams took me under his wing and helped me to develop as an information scientist. Our work together as teachers, researchers, and authors was always fun and always enjoyable.

[3] Teilhard de Chardin, Pierre (1964) The Future of Man. HarperCollins.

Organization of the Book

Chapter 2 provides a little bit of history. The history is generally accurate, but it is very selective in an effort to highlight what I see as the most important concepts and developments. Chapter 3 provides a basic definition of the illusive concept of information. Chapter 4 explains some of the things that make digital information so special. The chapters on information are important to understanding what is going on today. I have made every effort to minimize the mathematics and formal logic, but some reader may still find these a little dry. Chapter 5 looks at the evolution of digital technology. Chapter 6 addresses how documents are changing. Chapter 7 addresses the proposition that we are in the midst of a revolution and defines the nature of that revolution. The remaining chapters address some of the ways this revolution has already impacted our lives and more importantly lays out some of the important things we need to consider as individuals, workers and leaders in organizations, and as members of various social groups, from church goers to citizens of nations.

Chapter II: History

Introduction

It is always difficult to predict what the future holds, and given the amazing pace of the development of the digital world, predicting the future is particularly difficult. My first book on the future of documents was written in the late 80's and failed to account for the viral impact of the World Wide Web. The World Wide Web took us by storm. The emergence of e-books was long predicted, but it still arrived suddenly. Social networking is still emerging and morphing. This chapter provides a perspective on what has and is happening in our digital world with an eye to seeing if a look in the rearview mirror offers any clues as to a trajectory. The obvious conclusions at the end of this chapter are nowhere near as much fun as the process of trying to predict likely trajectories by looking at where we have been. We begin with the birth of telecommunications.

It is hard to pick a particular date or event that marks the beginning of this new era of immediacy. While there is no one date that is critical, I would suggest that we might define the date for the beginning of the era of immediacy as May 24, 1844, when Samuel F.B. Morse sent the famous words "What hath God wrought" from Baltimore to the Capitol Building along the wire. As I see it, there are two significant roots. The first has to do with the virtually instantaneous transmission of information across space – the electrical and later optical transmission of information. .The other has to do with creating a record that makes an event immediate to the consumer – an example would be a photograph. While the creator of the photograph imposes a perspective, it is much less than is imposed by describing what is viewed in words. We begin with early developments in both areas.

A Look Back at Photography and Telecommunications

Through the 1800's there were a number of experiments at producing photographs. In 1826, the Frenchman Joseph Niépce produced photographs on pewter plates covered with a petroleum derivative that hardened with exposure to light allowing the other areas to be washed away. The negative image can then be coated with ink and impressed upon paper, producing a print. Other developments included glass plates and other materials. In 1884 George Eastman developed a dry gel on paper, eliminating the needed for boxes of plates and toxic chemicals.

The use of a time series of photographs to create “moving images” was the natural next step in this process. In 1877, Eadweard Muybridge successfully photographed a horse in motion using a series of 24 stereoscopic cameras. The field was advanced significantly by William Dickson’s Kinetograph, which was developed in Edison’s Lab and patented in 1891. It took a series of photographs on standard Eastman Kodak photographic emulsion coated on to a transparent celluloid strip 35 mm wide. Dickson also developed a Kinetoscope that was designed for viewing by one individual. By 1894, Charles Jenkins had developed a projector that allowed viewing by a group – the Phantoscope. From this point forward, cameras and projectors were standardized on the 35-mm film and 16 frames per second. And the world of movies was born.

In parallel with the development of photography, work on electricity and electromagnetism was frenetic in the early 1800’s and important work began to crystallize in the early 1830’s. In this realm, there were three parallel developments -- the electrical transmission of codes, voice, and images. While there were some early electrochemical telegraphs and numerous experiments, it was in July, 1837 that William Cooke and Charles Wheatstone demonstrated a form of electromagnetic telegraph in London. In January 1838, Morse and his assistant Alfred Vail demonstrated a similar independently developed device near Morristown, New Jersey. The Morse device had some advantages over the British device in terms of distance and power. It also had the advantage of an optimum code, Morse Code, to aid transmission. The famous transmission between Baltimore and Washington – "What hath God wrought" – occurred on May 24, 1844. For the first time, it was possible to communicate virtually instantaneously over space. By 1861, there was a transcontinental telegraph cable in the US and by 1866 there was one across the Atlantic.

The telephone patent was applied for by Alexander Graham Bell in March 1876. This was the beginning of the technology for transmitting sound over wires. The next 25 years showed slow growth of the technology due to patent and business conflicts. By 1900, there were only 1,355,000 phones in the US, which had a population of 76,000,000. The numbers grew to 7,635,000 in 1910, 13,329,000 in 1920, 20 million in 1930, 22 million in 1940 and 43 million in 1950 when the population of the country was 152 million.

To the transmission of codes and sounds, we would add technology for the transmission of images. There were a number of individuals who developed machines capable of transmitting images electrically, but we select Ernest Hummel's 1898 *Telediagraph* as a commercially successful fax-like device for sending pictures via telegraph lines. In 1907, Edouard Belin invented a phototelegraphic apparatus, called the Bélinographe, capable of sending photographs to a remote location via telephone and telegraphic networks.

All of the developments above address transmission of information over wires – point to point. The use of the electromagnetic spectrum to transmit information was next. Radio communication was first demonstrated in 1893 by Nikola Tesla in a presentation before the National Electric Light Association. In December 1901, Guglielmo Marconi established wireless communication between St. John's, Newfoundland and Cornwall, England. In 1925, John Baird was able to demonstrate the transmission of moving pictures via what is now called mechanical television in London, England. The first version of an electronic television to show promise was produced by Philo Farnsworth of the United States, and it was demonstrated in Idaho in September 1927.

The Development of Distributed Computing

There are three technological developments that made huge contributions to the evolution of computing. They include networking, personal computing, and human centered computing. We discuss each of these briefly below.

Networking

The Internet has fascinating roots that most trace back to J.C.R. Licklider’s memos on man-machine communication. Licklider’s vision of a “Galactic Network” and the subsequent funding of the ARPANET by the Defense Advanced Research Projects Agency (DARPA), where Licklider was the first head of the computer research program, gave birth to the Internet we know today. From those first papers in 1962 through the 1960’s, progress was made on the idea with a number of people contributing. In 1968, Bob Taylor was able to convince his boss at the Advanced Projects Research Agency to provide funding to connect research projects across the country. The first message was from Leonard Klienrock’s laboratory at UCLA to Douglas Engelbart’s laboratory at SRI on October 29, 1969. While the full story is a most worthwhile read,[4] it is beyond the scope of the treatment here. This ARPAnet effort grew with time into the Internet—a system for addressing and messaging between machines. This led with time to an increasingly seamless computing environment that can connect every computer and computer enhanced device on the planet. From our point of view, the 1960's saw the birth of ubiquitous computing.

The vision of a world-wide network of interconnected machines was born a full thirty years before the web came about. There were controversies over protocols and approaches, but the basic vision was set. The parts of the vision included the notion of "packets" of information and a mechanism for moving them from machine to machine. Originally, like so many things related to the internet, the vision started small and the original Network Control Protocol (NCP) envisioned an address space of well under 1000 machines – up to 4 machines at up to 256 locations. The protocol was later refined and broken into two components. An Internet Protocol (IP) with a vastly expanded address space and a Transmission Control Protocol (TCP) that provided for sequencing packets and dealing with lost or damaged packets. The subsequent provision of a Domain Name Service allowed for human readable addresses that were translated by servers across the network in a distributed fashion. Finally, on top of this infrastructure of message passing peers, it was possible to construct a variety of other services – mail, file transfer, and remote connection.

While this networking scheme has been strained and continues to undergo modifications, it has proved dramatically stable and extensible. It provides the first of the three legs for the development of distributed computing.

Affordable Personal Computing

It is hard to imagine today that there was a time when computers were scarce resources. It is probably true that the use of microprocessors in devices of all kinds was inevitable. It may indeed be that the technology derives from the development of small computers for use in space exploration and calculators. Regardless, it is also true that IBM and the IBM PC, both because it was IBM's personal computer and because IBM provided a public specification for the system, were instrumental in moving

[4] Leiner, Barry, et. al A brief History of the Internet. Retrieved May 2011 from the Internet Society website: http://www.isoc.org/internet/history/brief.shtml

the micro computer into the front office and the offices of line operations. From there it was a much smaller step to the home.

We are getting very close to what I see as “cocktail napkin” computing, i.e., computing devices as accessible as a cocktail napkin or match book cover. When we can enter and retrieve digital information as easily as writing on a scrap of paper, we will have achieved ubiquitous commodity computing. The smart phone has made this very possible.

Accessible Computing

At the same time the IBM PC was being readied for the marketplace, Xerox, through its research at the Palo Alto Research Center (PARC) was getting ready to roll out the commercial version of its Alto, known as the Xerox STAR. The result of more than a decade of research, the STAR brought together a set of technologies in their full glory. This astounding leap was the result of a decision by Peter McCullogh of Xerox to fund research at a new research center in Palo Alto whose goal was the discovery of the “information architecture” for the office. This effort led to the object oriented windows, icons, mouse, and pointer based systems we use today. In the process, PARC researchers also gave birth to object oriented design and programming languages, the Ethernet, the laser printer, the bitmapped screen, WYSIWYG editing, etc. etc. etc. The mouse developed a decade earlier by Douglas Engelbart had been researched and refined into an ergonomic two button beauty. The chord device,[5] also pioneered by Engelbart, had been integrated in part into the mouse and in part to a set of universal commands represented by dedicated buttons on the keyboard. The “desktop metaphor” full blown and refined came forth from PARC along with the Ethernet, the bitmapped screen, the floppy drive, and the laser printer.

The richness of the environment put forward by Xerox is incredible even today. In some ways, the environment created at PARC has yet to be equaled by the two successful commercial systems based on it – the Apple Mac and the Microsoft Windows. There was great beauty and integrity of a fully object oriented working environment capable of WYSIWYG printing, intuitive controls, and virtual everything – keyboards, terminals, file systems, etc.

[5] The chord device was a way of issuing commands by pressing buttons. With five buttons, the device was capable of issuing 31different commands. Engelbart’s first mouse had no buttons for commands – it was only used to move the cursor.

Ubiquitous computing

The late Mark Weiser, a PARC researcher during the 1990's, proposed that the next generation of computing would be defined by ubiquitous computing. In the coming years, devices that are connected to the network and that have an intelligent control component will constitute a kind of invisible ubiquitous computing fabric that surrounds us. These devices will have data acquisition mechanisms and narrowly scoped intelligence that will allow them to work with us. The key to the next generation devices is that they will be special purpose, not general purpose, and they will for the most part have invisible interfaces. The car is talked about as one of these devices. Today's car already has a significant amount of computational power. Connected to the internet it will be able to deliver information about traffic, lodging, restaurants, etc. Refrigerator's will have bar code readers and communications capability to order needed goods. Thermostats will sense human usage patterns making it unnecessary for human to adjust or set the thermostat. E-business will be busy over the coming decades finding ways to package and sell just-in-time information to consumers who will be on the move.

A Selective History of Computers and Networking

It is of some interest to reflect on the fact that a punch card machine developed by Herman Hollerith was used in 1889 to help count census information. It was this device, along with a meat scale and a time clock that were later combined to form the International Business Machine (IBM) Corporation. While this early history is of some interest, the formal history of computing and networking is the focus here. After a brief review of key computing events, we look at the development of the internet, and finally review the explosion of the World Wide Web. The last part of the chapter looks at the magnitude of the changes over the last 50 years.

Looking at the broad selective timeline of computing and networking events presented in Figure 1, there are some events that might be highlighted. While we are a full decade into the second half of the first century of computing, there is still some utility in reflecting on the difference between the first and second halves of the century. I observe three things using my fractured timeline.

First, the Web is barely 20 years old, and for most people 10-15 might be more accurate. Given the age of the Web it is more likely that it has yet to develop a final form than computing and networking. Second, while few things stand the test of time in this realm, the "desktop" is a notable exception. C. Peter McCologh,President of Xerox charged the first Director of PARC, George Pake with inventing "the information architecture of the office." Out of that challenge came many things, but

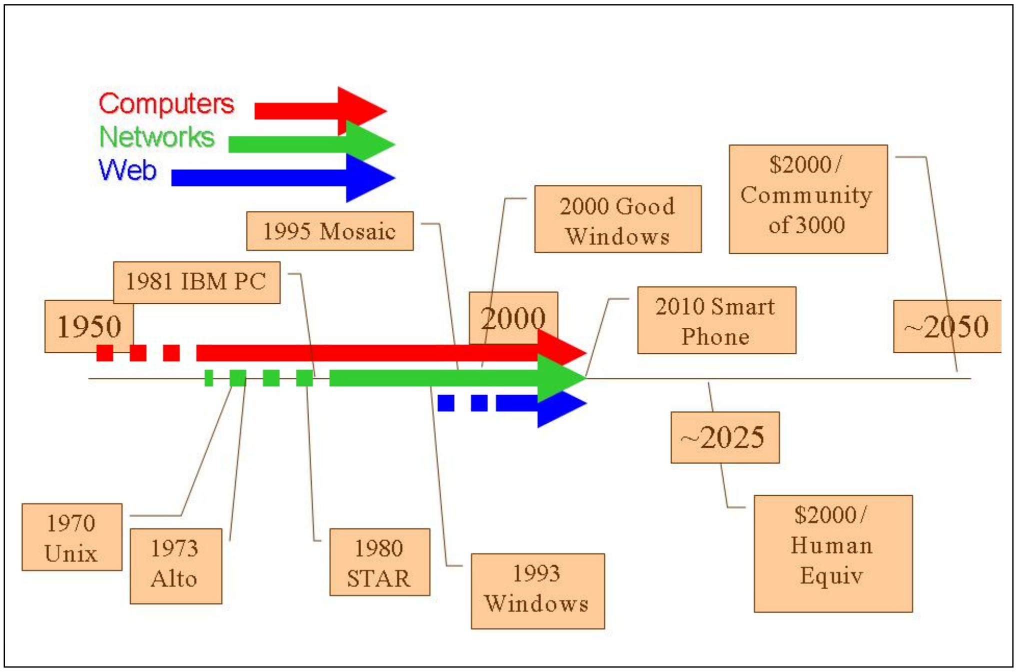


Figure 1: Evolution of Computers, Networks, and the Web

foremost was the "Alto" and the subsequent "Star" personal computers. And of all the innovations made manifest through these machines, a case can be made for the greatest being the WIMP (windows, icon, mouse, pointer) interface and the "desktop metaphor." This is the same desktop that you are sitting at today – more than 30 years later. We use programs hung on this metaphor to process information – to communicate and collaborate. Third, while predictions of exact trajectories vary, there is still solid support for Moore's law – a continued doubling of computing power every 18 months or so.

Every time I use my Droid X, marvel at the 1 Gigahertz device in my pocket that "doubles" as a phone and is always connected via four different networking technologies – Wi-Fi, 3G, Bluetooth, and GPS. In 2025, some suggest the personal device, which will likely rely on the grid, will have the storage capacity and processing capability of a single human. Around 2050, the optimists suggest it could have the capabilities of a community of 3000 humans. This should not be taken too literally. It is much more likely that my personal device will operate a series of agents operating around the clock on my behalf. Even if they are all idiot savants, it is more than likely that they will be capable of managing my calendar, knowing my travel preferences, keeping track of what I write, doing basic searches for me, etc. (I find it hard to imagine any way that I will be able to keep 3000 specialized assistants busy.) So, the broad picture is one of immensely powerful personal assistants working to help me communicate, collaborate, coordinate, and calculate in such a way that they operate beyond the vision of Vannevar Bush as an intimate supplement to human memory, and dare I say, intellectual activity. (Bush's vision is described in detail on page 85.)

The Evolution of the Internet

The growth and development of the internet is more difficult to track. The theory behind a packet switched network was laid out in the 1960s. The first nodes of what would become the internet were actually put in place as part of Arpanet beginning in 1969. The protocols that would become essential to the development of the packet switched network were developed through the 1970s. A protocol is a set of rules that define how communication between machines should take place. The protocols developed include telnet (remote connection) in 1972, FTP (file transfer) in 1973, NCP (network flow control) in 1974, and SMTP (e-mail) in 1977. NCP was replaced by TCP/IP in 1982. (The Internet Protocol, IP, describes how to address messages and the Transport Control Protocol, TCP, describes how to insure messages are delivered reliably) DNS (alternative naming for addresses) was introduced in 1984. The World Wide Web (WWW) was released by CERN in 1991 and took off with the first graphic browser – Mosaic – in 1993.

During the 1990s, things changed rapidly. We forget now about gopher and newsgroups, but they had their day. Early on file transfer dominated in terms of packets on the internet, with telnet connection packets being second, and e-mail being third. It is a little hard to trace the

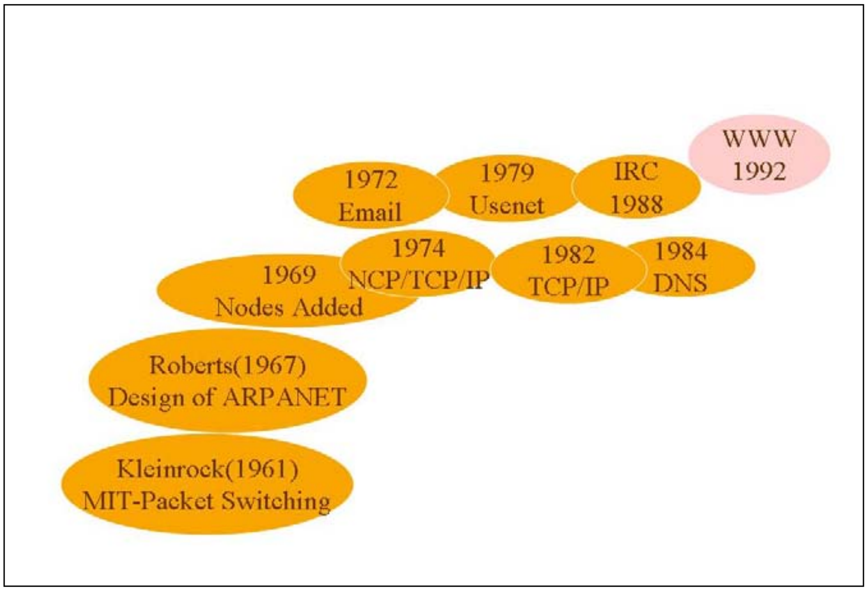


Figure 2: Development of Internet Protocols

growth of e-mail, but the phenomenal growth of the World Wide Web (WWW) is clearer. Web (http) traffic exceeded telnet packets in 1994 and ftp in 1995.

Andrew Odlyzko's 2003 paper on "Internet Traffic Growth" provides a balanced perspective on internet growth. He estimates that from 1990-1994 traffic about doubled every year from 1.0TB in 1990 to 16.3TB in 1994. Data for 1995 was not available, but by 1996 it had risen to 1500TB – the WWW explosion. He documents an ongoing growth rate of near 100% per year from 1996 through 2002 – from 1500TB to around 100,000TB. Obviously, both the WWW and P2P protocols were largely responsible for this growth. Further analysis shows SMTP traffic (estimated based on partial data) to have doubled about 4.5 times in 10 years. During roughly the same time, http traffic doubled 9 times! In 1994, e-mail traffic accounted for 16% of internet traffic and the web accounted for virtually none. By 2003, e-mail traffic had dropped to less than 2% and web traffic had increased to almost 50%. The number of e-mail and web

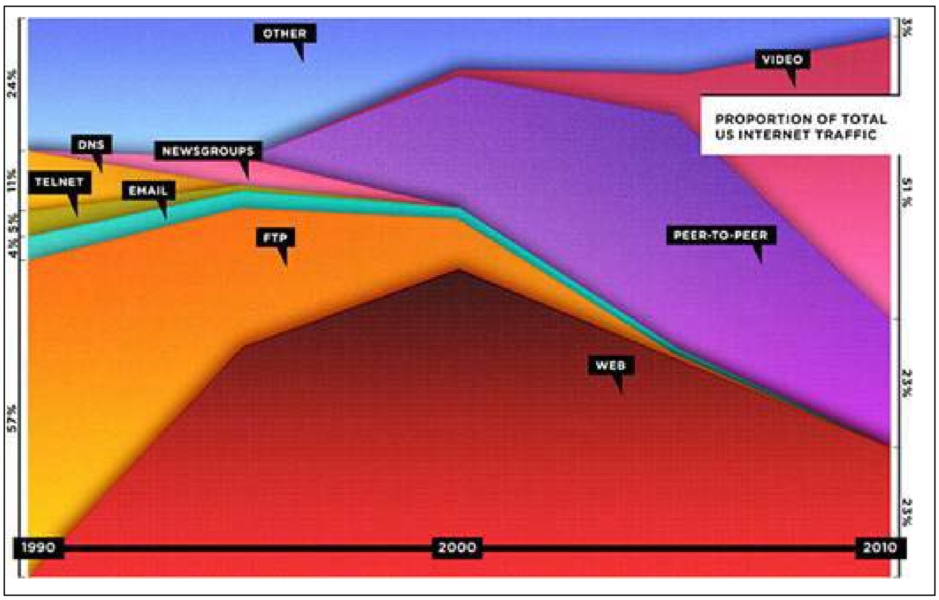


Figure 3: Breakdown of Network Packets by Type

packets moved continues to increase at a rather amazing rate, i.e. there are more http packets this year than last. At the same time, given the totality of the bandwidth available, new applications are absorbing a larger percentage. Figure 3, published in wired magazine[6] is based on data from the

[6] Chris Anderson, Chris & Wolff , Michael, (2010, August 17) The Web Is Dead. Long Live the Internet. Wired Magazine. Retrieved May 2011 from Wired website: http://www.wired.com/magazine/2010/08/ff_webrip/.

Cooperative Association for Internet Data Analysis presents an interesting overview of the trends.

What does this suggest about the future? It would seem reasonable to suggest that we will continue to find new uses for the internet, which continues to grow at an astonishing rate. This network bandwidth has allowed us to begin to embrace a network centric model of computing. The desktop no longer needs to be a storage facility. Streaming gigabytes of data to watch a movie has become cheaper than mailing a DVD or driving to a store to pick it up. Music and video packets now crowd the internet. Communication and entertainment applications absorb a growing percentage of the bandwidth with video alone accounting for more than half of all internet traffic in the US! While the number of packets devoted to the WWW continues to grow in absolute numbers, they have begun to decrease as a percentage of the total packets.

The Emergence of the World Wide Web

After thirty years of slow growth and development, the World Wide Web emerged on a connected base(ARPAnet) of affordable computers(IBM PCs) which most people could easily learn to use (Xerox Star). Equally as important, by 1995, most people had accepted as trustworthy digital information technology they used daily in the form of ATM machines and supermarket checkout scanners. Given this environment, the very simple "innovation" called the world wide web caught on like wild fire. (I put innovation in quotes because the basic concepts of the web—a client server protocol, a hypertext form of organization, and a structured document format for data interchange were 25, 40, and 10 years old respectively at the time they were put forward.) This is not to diminish in any way the import of what Tim Berners-Lee did at CERN. As is the case in all revolutionary technologies, both the environment for the technology innovation, and the development of the technology itself are always found to have a longer and deeper history than the critical event that caused the revolution. Now we are at the millennium. E-business is here and will play a role in the coming decades reshaping how we do business. How do we get a handle on what is going on?

Very few killer technologies arrive as a surprise. While the World Wide Web is no exception to this rule, it did move into the forefront of our experience more rapidly than most. It is almost as if the 30 years of development of the internet were waiting for the right use – a use provided by the World Wide Web. Most of us became aware of the Web in 1993 when the Mosaic browser took hold on Unix workstations running X Window System graphical interface. For the record, some of the key early events leading to the explosion included:

March, 1989	Tim Berners-Lee (TBL) working at the CERN, wrote "Information Management: A Proposal" and asked for comments
May, 1990	TBL recalculated the proposal
Sept,, 1990	TBL is given permission to explore the concept of a global hypertext system.
Oct., 1990	TBL starts work on a hypertext GUI browser+editor using a Next Machine TBL coins the term WWW
Nov., 1990	TBL develops a WYSIWYG browser/editor with direct inline creation of links.
May. 1991	Presentation to "C5" Committee. General release of WWW on central ERN machines.
Aug., 1991	Files posted to the internet -- ftp
Dec., 1991	Demonstration at Hypertext'91 in San Antonio, Texas
Jan., 1992	Line mode browser release 1.1 available by anonymous FTP
May, 1992	Pei Wei's "Viola" GUI browser for X.
July, 1992	Distribution of WWW through CernLib, including Viola.
Jan., 1993	Midas (Tony Johnson, SLAC), Erwise (HUT), and Viola (Pei Wei, O'Reilly Associates) browsers +Mac browser (ECP). Around 50 known HTTP servers.
March, 1993	WWW traffic measures 0.1% of NSF backbone traffic.
July, 1993	NCSA httpd prototype made available
Sept., 1993	Marc Andreessen releases "Mosaic for X."
Sept., 1993	WWW traffic measures 1% of NSF backbone traffic1993, Over 200 known HTTP servers.
March, 1994	Marc Andreessen and colleagues leave NCSA to form Netscape.
May, 1994	NCSA httpd 1.3
July, 1994	MIT/CERN agreement to start W3 Organization is announced by Bangemann in Boston.
Dec., 1994	First W3 Consortium Meeting at M.I.T. in Cambridge (USA).
April, 1995	First public release of the Apache web server.

By 1995, the web had taken off. The Mac and PC were both evolving graphical displays that made simplistic use of the web possible. Looking at the graphic below, one cannot help but be impressed with both the growth and dormancy of the web. The technology and the infrastructure were in place by 1994, but it was not until what we call the social web took hold that real explosion occurred.

Figure 4 shows a couple different things. First, all of the technologies in yellow are non-web technologies. While most are internet technologies, I have included other seminal developments that have in one way or another influenced the development of the internet. (Internet Relay

Chat, now famous as a control mechanism for bot-nets has been with us as a precursor of social networks since 1988! Instant Messaging (IM) and the Short Message Service (SMS) go back to the early 1990's.) The web gave birth to websites, links, tags, and wikis. These base capabilities were subsequently combined (or mashed-up) into Friendster, Wikipedia, and a variety of other collaborative efforts. The traditional web is being replaced by the social web. We spend relatively less "web time" searching for information (Google) and relatively more time sharing with friends (Facebook).

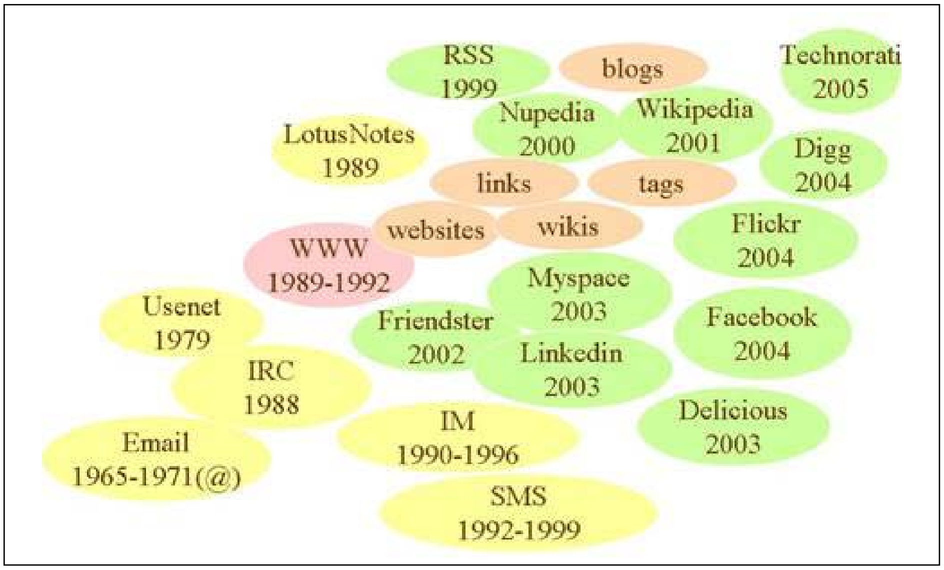


Figure 4: Evolution of Web and Pre-Web Technologies

I am reluctant to say where the web is heading, but it would seem fair to suggest three simultaneous trends that will grow in a complimentary fashion. First, the web as a vast information store is not going to go away. It will continue to grow. While not as evident, the "semantic web" is slowly taking form – consider the special sections of structured information available in Wikipedia. Finally, it is very clear that a new "social web" is growing by leaps and bounds. This web provides rich new venues for communication and collaboration.

The Magnitude of Change

One way to think about the changes that are occurring today is to measure the magnitude of the changes. Many are familiar with Moore's law, which suggests a doubling of computer power every eighteen months given constant cost. It is interesting to think about the changes in storage, transmission and processing power during the first 50 years of the digital era. These increases are large enough to measure in orders of magnitude.

That is, like the Richter scale that used to be used for earthquakes, each number represents an order of ten increase over the previous number. A magnitude of one represents a ten-fold increase. A magnitude of 2 equals a hundred fold increase, etc. Historically, a Richter scale earthquake of magnitude 7 or 8 was devastating.

Another way to think about this history is to make an assessment of the extent of the changes, the "infoquakes" that have taken place in information storage, processing, and transmission over the last 50 years.

Processing

Searching for information is an important process. Today, documents that would fill a library can be searched electronically in minutes. The electronic search can locate information within documents that could not be located by humans in a physical library in any reasonable amount of time. For example, any search that could not be answered based on bibliographic data and abstracts would require reading every document and would be impossible for all practical purposes. This human processing limit drives the need for bibliographic information. Electronic search allows for the storage and retrieval of full text, reducing the need for the analytic surrogates. The United States Library of Congress contains about twenty million books. Any word could be located by doing a binary search of an index of these books and such a search could be completed within 42 reads of the index.[7] This means the search would be completed in a few milliseconds. This does not imply that the catalogers, indexers and librarians are not needed. It does imply a change in their role. Skilled indexers will still be needed to add index terms that do not actually appear within the text. Similarly, a 1965 concordance of Byron's works took 25 years to develop manually.[8] The 285,000 entries can be generated by a computer in a matter of minutes. While concordance construction will change dramatically, expertise will still be needed in the use of concordances. Looking at the quake in processing capability, one can make a variety of measurements. Overall, the magnitude of the change in information processing speed via technology between 1950 and the year 2010 is in the range of a magnitude 7 infoquake. While the visible processes are no longer increasing dramatically, there are many more helper processes working on our personal machines to support us. Two examples would be mobility functions and speech processing.

---

[7] Witten, Ian H., Moffat, Alistair and Bell, Timothy C., (1994) **Managing Gigabytes**, New York: Van Nostrand Reinhold, p. 15

[8] Witten, Ian H., Moffat, Alistair and Bell, Timothy C., (1994) **Managing Gigabytes**, New York: Van Nostrand Reinhold, p. 2

Storage

Storage media are experiencing orders of magnitude shocks. The density of memory on silicon chips keeps increasing. The first PCs in the early 1980s contained chips with 16K bits on a chip. PC memory chips now hold 1 Gigabit. State of the art chips are now being developed that store 100 GigaBytes in a cubic inch. This corresponds to a increase of more than a 100,000 fold. Personal computer hard disks have gone from being nonexistent to having terabyte disks readily available. Similarly, removable media have gone from 360,000 byte floppies to CD-ROMs holding about 650 MegaBytes, DVD storage media with a capacity of 4.7 GigaBytes, and now Blu-ray with a storage capacity of 50GB. This represents an increase of 140,000 times. Looking at the quake in storage capability, measurements would seem to suggest that from the 1950 to the year 2010 the magnitude of the change in information storage capability via technology will be in the range of a magnitude 7 infoquake. It is hard to believe that the developments are showing no signs of slowing down. My most recent PC has no hard drive – storage is solid state memory of incredible speed. Smart phones, music players and thumb drives offer us Gigabytes of storage at a retail cost of about $1/GigaByte!

Transmission

Transmission capacity represents another infoquake. Teletype machines ran at 150 or 300 bits per second. By 2000, 28Kbps modems were common worldwide. That is two orders of magnitude over teletype. Digital Subscriber Lines (DSL) offer a 4-20 fold increase providing speeds of up to 3 MegaBits/second. Broadband services provide a variety of different options for connection speed, but it is not unusual to see rates of 7-10 MegaBits/second. This is significant when one considers that to transmit a book with 210 pages, ten images and 20 graphics would take a little more than an hour and a half by 2400 bps modem but less than 5 seconds via a 3 Megabit connection.[9] Broadband can provide serious users with the 10,000,000- 100,000,000 bps network connections many academics currently have on their desktops. In this context, transmission capacity from 1950 to the year 2010 the magnitude of the change in information transmission capability via technology will be in the range of a magnitude 8-9 infoquake. While the transmission infoquake is significant enough in its own right, the co-occurrence of the three infoquakes is truly unprecedented.

[9] The calculations are based on a book consisting of 210 text pages each with 2400 characters of text, 10 one Megabit images and 20 graphics of 16,000 bytes.

Conclusion

In 1998, Denning and Metcalfe, edited a book entitled **Beyond Calculation** to celebrate the 50th anniversary of computing. The title provides a cornucopia of overlapping meanings. The most interesting is the observation, also made by several contributors, that the era of computers that calculate is giving way to an era in which calculation is giving way to communication, collaboration, and coordination. We are in a period of significant experimentation and we are finding uses beyond calculation for this disruptive technology.

When one thinks about experimentation with new technologies, several thoughts come to mind. One is that important new technologies are often “less” technologies. The automobile began as a horse-less carriage. Cell phones first emerged as cordless phones. We still talk about wireless networks, and paperless offices. Historians of technology point out that this perspective on new technologies often prevents us from embracing the full potential of the innovation.

Incunabula is a term used by librarians to refer to books produced between 1453 and 1500 – the first half century of the disruptive technology of moveable type mass production printing. Incunabula books are remarkable in the degree to which they differ from each other and what we think of as books today. They were experiments. Indeed, the term incunabula is more broadly defined as any art or industry in the early stages of development. So we may think of what happened during the first half century of computing as computing in its early stages with lots of experimentation.

Chapter III: Information

Introduction

This book is about digital information. We all know what information is – that is, until we sit down to define it precisely. And before we can begin to come to grips with digital information, it is useful to come to grips with what we mean by information generally.[10] Unlike other fields, information science lacks clear concise definitions of what it is about. What follows provides some bounds on the concept of information. There is no unified theory of information. At the same time, what follows should allow a more informed discussion of this difficult concept. We begin with what most of us already know and continue on to think about information from a variety of perspectives to conclude with ways to measure information.

The common sense understanding of information is that it increases what we know. If I tell you that it is raining, I have given you information. If you already know that it is raining, I have not given you any information. Note that information is a bi-product of communication – a message – and that the same message can be information to one recipient and not to another. The more formal definition for information came from research on telecommunications channels at Bell Labs. World War II saw significant intellectual resources devoted to cryptographic analysis – code breaking. Claude Shannon joined the cryptography effort at Bell labs and much of that work led to his later publications on communications theory. Put in the simplest terms Shannon suggested that the greater the uncertainty (entropy) in a message, the more information required to transmit the outcome. Again, at the risk of oversimplification, communicating the outcome of some event where there are a dozen possible outcomes required more

[10] Curiously, my research area in "information" science is structured documents. Just like information, we all know what documents are, that is until we attempt to come up with a precise definition. In Chapter VI:, we will take a closer look at how we define documents.

information than an event with only two possible outcomes. From a digital information perspective, using a binary number system, we can identify two outcomes using one bit – 0 or 1. Again using only zero's and ones, binary numbers, an event with four possible outcomes can be represented by the binary numbers 00, 01, 10, and 11.

A Common Sense Understanding

Perhaps the best definition of information is based on the common sense usage of the term. Basically, if you tell me something I don't know, you have given me information. If I already know it, you have not given me any information. If I tell you that it is raining outside. I have given you some information if you did not know it was raining. If you knew it was raining, we would say that no information has been provided. This is very close to what we posit on a daily basis when we talk about the human process of providing information.

Information as a Human Phenomenon

Can information be separated from the human experience? Can you inform a machine? Can you inform a dog? Some people are going to say yes, some no. Does a book contain information? Some people are going to say yes, some will say no. I begin with the premise that in a world without humans, there is no information. This proposition is similar to the question about whether a tree that falls in the forest makes noise if there is no human to hear it. While I am fully prepared to extend the concept of information beyond humans, there is some benefit in beginning with the premise that information is tied to human experience.

For a human, a message conveys information, if it tells us something we don't know. This suggests that information is relative. A message delivered to two people varies in the amount of information it carries based on the state of the knowledge in the receiver. In addition, getting information changes us – we "know" something more. The same message delivered to a person twice might contain information when first delivered, and no information the second time it is delivered. These common sense observations suggest four things:

1. Information is carried (in a message)
2. Information is not absolute, but relative for each receiver
3. Information can have a magnitude (at least it can be zero or more)
4. The amount of information is influenced by what the receiver knows

The rest of this chapter takes a variety of perspectives on information. First, we explore the context that determines whether a message contains information. Second, we look at the difference between oral and symbolic communication of information. Third, we explore the measurement of the magnitude of information. Fourth, we turn to a related discussion of noise. Fourth, we think a little bit about treating information as an object – a commodity. Fifth, we look at the antithesis of information – noise. Finally, we look at thinking about information as a process.

Information in Context

Many readers will have already concluded that how we as humans deal with information depends on what we know – is there a relationship between information and knowledge? Again, I would posit that both terms are most meaningful in the context of the human experience. In information science, it is not uncommon to see a diagram such as this.

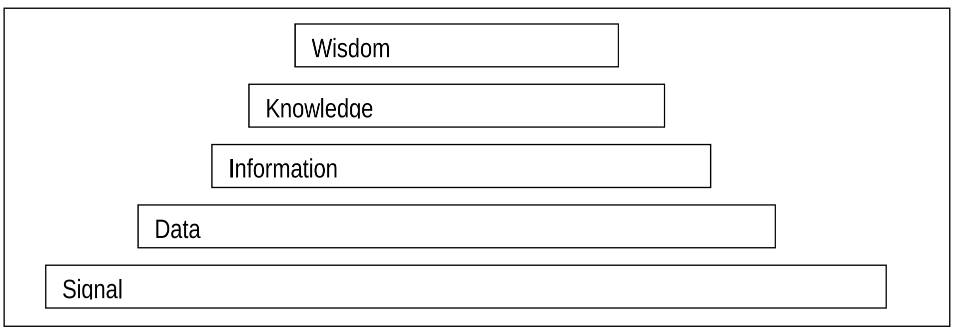


Figure 5: Traditional Information Hierarchy

This pyramid implies a relationship between these five terms. In most models, signals are the raw material and data is built from signals. Similarly, information is distilled from data, etc. While there is some validity to what is being implied, the model is not particularly satisfying. A

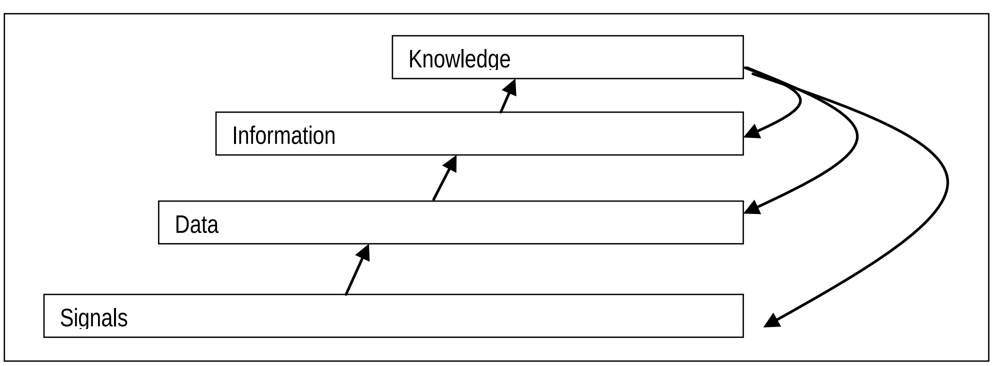


Figure 6: Impact of Knowledge on Information

better picture emerges when a couple changes are made to the diagram.

First, let's postulate that a variety of events in the universe generate signals. We suggest, again for purposes of argument, that signals move through space and time whether human exist or not. Thus, related to the tree falling in the woods, I would suggest that sound waves move through the forest when a tree falls, but those sound waves are not judged as data that a tree is falling if there is not someone or something to hear them. Second, not only information, but data, knowledge, and wisdom are aspects of the human experience; that is, without humans, it makes little sense to talk about them. We are misled by the notion that there are commodities called data, information and knowledge apart from the processes by which humans, recognize, get informed, and know. The simplistic pyramid provided in Figure 5 then becomes something like Figure 6 which shows the critical importance of knowledge on the interpretation of signals, data, and information. We have already discussed the fact that what we know affects whether or not we see messages as containing information. Similarly, what we know impacts whether or not we recognize signals as data or organize data to create information.

But our exploration is still incomplete. Is there anything more than knowledge? How is something like wisdom related to our structure? When what we know allows us to solve problems in the real world, we might expand knowing to knowing how. Wisdom is the process by which we apply what we know to the world around us. While it may be overly simplistic, knowing many things does not assure that we will use the information correctly. Knowing what knowledge and information is relevant to the solution of a problem is important. This category of intellectual activity may be operationally defined as wisdom.

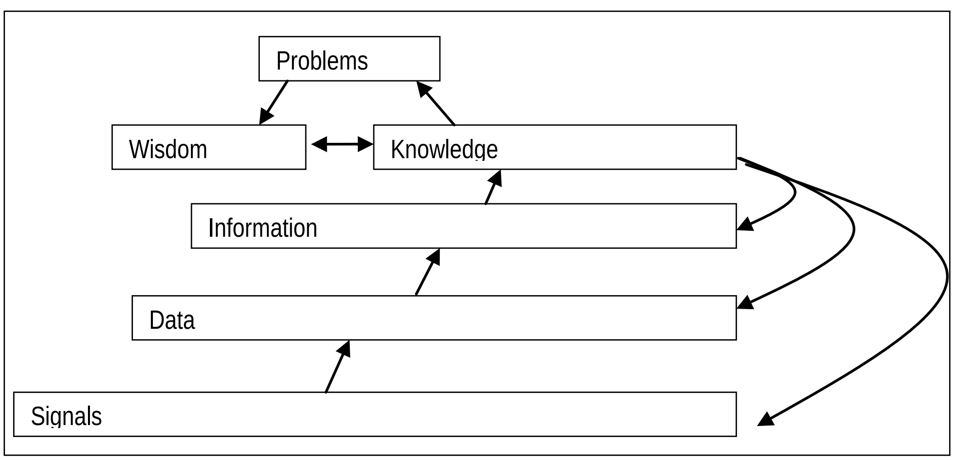


Figure 7: Knowledge and Wisdom

In summary, information is intimately tied to the human experience. What we consider information is influenced by what we know. And what we know is determined by information we acquire. Information may be acquired from data which in turn may be constructed from signals. Both of these transformations are ultimately tied to the knowledge we use to conduct the transformations. Finally, what we think of as wisdom relates to the correct use of knowledge to solve problems. Wisdom is the application of knowledge to problem solving.

Communication: Oral and Symbolic

This picture we have drawn is somewhat satisfying, but it is not an accurate depiction of where we are. What we implied in the first few sections and then ignored in the section on context is the import of communication in transferring information. Indeed, what we have suggested is a rather sterile process by which signals are organized into data which is them further organized so as to constitute information.

Chapter 7 will propose that a new form of communication is emerging. To appreciate this latest development in the communication of information, which is most often exchanged in communications made up of symbolic messages, there is some value in thinking about how communications have changed over the years. Keep in mind that there was a time when humans communicated with grunts and gestures.

Humans distinguished themselves from other species in the ability to communicate using a sophisticated language. Discussing this development frequently leads to arguments that other species communicate as well, and that is not something with which I would disagree. At the same time, it is very clear to me that humans communicate via a spoken language that is quantitatively and qualitatively different from the language of all other species.

What is amazing is the relatively brief period of history during which we have used language to communicate. The fossil evidence seems to suggest that modern man, as evidenced by bipedal locomotion began to emerge two to four million years ago. In contrast, evidence of structures that would support modern speech doesn’t appear in the record until about 300,000 years ago. Our ancestors of 200,000 years ago were pretty much anatomically like us. Experts in this field seem to be in general agreement that the evidence suggest that language developed 50,000 – 100,000 years ago. My goal is not to question the judgments of the experts or to conclude that the story they tell will not continue to evolve. It is simply to observe that our species has walked upright and made tools for 2.5-2.6 million years or more. In contrast, we have been using a sophisticated language – telling stories and communicating efficient messages containing information – for less than 100,000 years, possibly for as little as 50,000 years.

The significance of this development is a matter for experts to debate and refine. What should be clear is that when human began using language made great leaps in the development of technology. Farming, herding, and tool development all escalated at the same time we developed the ability to communicate information via language. The reflection on what this social system must have been like is a fertile ground for thought experiments. For example, prior to the development of language, we might imagine that the limit on the knowledge and wisdom an individual might accumulate would be bound by their ability to observe and gather signals and data for themselves. Then the limit changed and became bound by the amount of knowledge that could be communicated by language between adult and child. The acquisition of new knowledge from the environment was still important, but memory became equally important as did the devices for passing on that information and knowledge in ways that made it easy to maintain.

The rate at which our species developed changed significantly with the development of language. While impressive, a more significant development was still on the horizon. Cave drawings have been found dating back as far as 30,000 – 35,000 years. Somewhere around 7,000-8,000 years ago systems of writing began to emerge where the symbols were glyphs representing objects, or ideas, or words. The evolution of pictographs into well developed writing systems such as Egyptian hieroglyphics occurred about 5,000 years ago. What we consider more modern writing systems where the symbols represented consonants and vowels occurred about 4,000 years ago. The classification of the many developments is more than can be covered here, but the point is that a comprehensive and flexible system for writing what we were able to say has been with us for about only 4,000 years. That is, we have been speaking for a small part of the time we have been making tools, and writing for only a tiny fraction of the time. Using the metaphor of a 24 hour day, of the 24 hours that we have been making tools and walking upright, we have been using spoken language for a little less than an hour, and have been writing down our thoughts for less than two and a half minutes. If we accept the fact that the development of mass production printing did much to speed up the spread of knowledge and information, it is humbling to realize those 650 years represent only the last 23 seconds of our 24 hour day.

Put simply, language allowed us to more efficiently transfer information and knowledge between individuals. That capability had a dramatic impact, but it still required the individuals exchanging information to be co-located in space and time. With the development of writing, the requirements of collocation and contemporaneousness for the exchange of information were eliminated. We were still limited in the amount of information and knowledge that we could absorb in a lifetime, but compared to the one to one exchange of information stored only in wetware, we virtually eliminated all the barriers.

Every time we talk about vast increases in the amount of information we can exchange, it begs the question of whether or not we can somehow measure information or knowledge. We turn to this topic next.

Measuring Information

Shannon's mathematical model of communication is often pointed to when one asks for a definition of information. They were defining the capacity of a communications channel and analyzed that capacity under the constraint of transmission error.[11] In 1948, working at Bell Labs, Shannon was thinking about noisy telephone systems. The question that Shannon was asking related to the capacity of a channel in the face of noise. He began with the proposition that the amount of information could be measured by the amount of variability in the message. The entropy of the message – the degree of uncertainty – is a measure of the information content. The analysis describes the amount of information that can be transmitted across a channel, given the transmission errors of that channel. An information source generates a message which is encoded as a set of signals sent across a communications channel which is subject to noise. The basic model they were using may be depicted as follows:

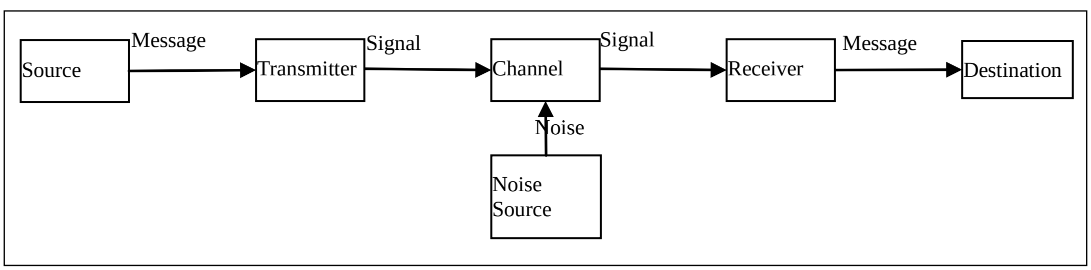


Figure 8: Signal and Noise in a Communication Channel

[11] Shannon, Claude (1948). "A Mathematical Theory of Communication". *Bell System Technical Journal* **27** (July and October): pp. 379–423, 623–656.
Shannon, Claude & Weaver, Warren, (1963). *The Mathematical Theory of Communication*, Urbana, IL: University of Illinois Press.

Probabilities, Entropy, and Information

People are often confused by the formulas we are going to show below, but they are easier to understand if we begin simply and intuitively. The first thing we need to do is get a sense of what we want the formula to tell us. Let's begin with figuring out the amount of information in a message that tells you the result of a coin toss. If the coin is rigged such that it always comes up heads, telling you it came up heads would provide no information. On the other hand, if it was a fair coin, which comes up heads half the time and tails half the time, telling you the result would give you some information. If we imagine rolling a fair die, telling you it came up "1" would give you some information because it could have provided any of the six numbers. The key insight here is that the probability of the message being one thing versus another is also a measure of the amount of information in a message that reveals the outcome. The probability of a outcome that is certain is always 1. The less sure we are of the outcome – the more possible outcomes there are – the smaller the probability of an outcome.

The next thing we need to do is turn to a little high school math. Familiarity with exponents, logarithms, and number systems are parts of math most of us have long since forgotten. For our purposes, we simply need to recollect that logarithms are the "powers" to which the base of a number system is raised to produce a given number. Working in our natural number system – base 10 – we produce the number 10 by raising 10 to the first power –the logarithm of 10 is 1. We can generate the number 100 by raising the number 10 to the second power – the logarithm of 100 is 2. Similarly, the logarithm of 1000 is 3. The logarithm of 42 is 1.623249. Don't worry, there is only a little more we need to remember. First, the logarithm of 1 is 0. It is actually true that the logarithm of 1 is always 0, regardless of the number system. The other thing we need to keep in mind is that the logarithms of numbers less than 1 – fractions – are negative numbers. Hoping those of you who have forgotten your math don't get too stressed out, the logarithm of 1/10 is -1 and the logarithm of 1/100 is -2. There are logs for every positive number – the logarithm for 1/5 is -.069897.

Using these properties of probabilities and logarithms, the amount of information in a message might be calculated the following simplification of Shannon's formula. He made use of one of the properties of logarithms we just discussed to produce an interesting result – namely that the logarithm of 1 is 0. The formula says that the amount of entropy, or information, in a message is the inverse of the sum of the probabilities of the various possible outcomes times the log of the probability. (We use the inverse – negative – because we want our numbers to be positive, and the logarithms of probabilities which are all greater than 0 and less than or equal to 1 are negative numbers.)

$$I = -\sum_{i=1}^{n} p_i \log_{10}(p_i)$$

Let's look at our coin toss again. If the coin is fair and it comes up heads 50% of the time and tails 50% of the time, the probability of heads is .5 and the probability of tails is point .5. This gives us the amount of information as:

I = -1*(.5 * log(.5)+ .5 * log(.5))

I = 0.3013

In contrast, if there is only one outcome, with a probability of 1, the formula yields the following:

I = -1 *(1.0 * log(1.0))

I = 0

This math, which proves quite powerful, is exactly what Shannon wanted. However, to get to his formula, we need to make one more change. Shannon actually defined the amount of information as the sum of probabilities of the outcomes for each component <u>times the base two log</u>.

$$I = -\sum_{i=1}^{n} p_i \log_{2}(p_i)$$

Using this formulation, a rigged coin still produces an information measure of 0. However, the fair coin produces an information value of 1. If there are four possible outcomes, each with a probability of .25, the formula generates a value of 2.

I = -1 *(.25 * $\log_2$(.25) + .25 * $\log_2$(.25) + .25 * $\log_2$(.25) + .25 * $\log_2$(.25))

I = -1 * (-.5 +-.5+-.5+-.5)

I = 2

This has some significant implications for computer science in that the numbers produced relate to the number of binary digits or bits required to transmit or store the message in a computer. Put more simply, if there are four possible outcomes, this equation indicates that the information can be communicated in two bits – outcome one would be transmitted as "00", outcome two as "01", outcome three as "10" and outcome four as "11." Using this same formulation we can determine all sorts of things. For example, when there are many possible outcomes with varying probabilities that are not even, the formula suggests that the number of bits of information needed should decrease.

This formula tells us a lot. For example, consider the amount of information in some word. If we consider the alphabet of 26 letters has co-equal outcomes. Shannon's formula would yield the following amount of information in a string of 26 letters.

I = -1 *((1/26*log2(1/26)+1/26*log2(26)…)

I = 4.70044 bits

However, as we all know, letters in the English language don't occur with equal frequency. This suggest that the amount of information in the alphabet will be different. The frequency of letter occurrence in English is as follows:

a	8.167%	j	0.153%	s	6.327%
b	1.492%	k	0.772%	t	9.056%
c	2.782%	l	4.025%	u	2.758%
d	4.253%	m	2.406%	v	0.978%
e	12.70%	n	6.749%	w	2.360%
f	2.228%	o	7.507%	x	0.150%
g	2.015%	p	1.929%	y	1.974%
h	6.094%	q	0.095%	z	0.074%
i	6.966%	r	5.987%		

Note that the "e" occurs much more frequently than a "q" or a "z." An 'e' has a greater probability of occurrence than a 'z'. When the formula is applied to this set of probabilities, we discover:

I = -1 *(.08167*log2(.08167) + .1492 * log2(.1492)...)[12]

I = 4.17576 bits

This indicates that we would need 12% fewer bits to store information based on these probabilities, which may not seem like a lot,

[12] The letters of the alphabet their probabilities and the logs of those probabilities are:

Letter	Prob	Log	Letter	Prob	Log
a	0.08167	-0.29516	b	0.01492	-0.09051
c	0.02782	-0.14377	d	0.04253	-0.19374
e	0.12702	-0.37812	f	0.02228	-0.12228
g	0.02015	-0.11351	h	0.06094	-0.24598
i	0.06966	-0.26774	j	0.00153	-0.01431
k	0.00772	-0.05417	l	0.04025	-0.18655
m	0.02406	-0.12938	n	0.06749	-0.26248
o	0.07507	-0.28043	p	0.01929	-0.10988
q	0.00095	-0.00954	r	0.05987	-0.24319
s	0.06327	-0.25196	t	0.09056	-0.31379
u	0.02758	-0.14287	v	0.00978	-0.06529
w	0.0236	-0.12756	x	0.0015	-0.01407
y	0.01974	-0.11178	z	0.00074	-0.0077

unless you are paying for extra sets of copper wires to connect phone central offices, or spending millions on storage facilities. The details of how finding like this are used can be very complex. Technically, one piece of the result is the development of variable length codes – called Huffman codes. While Huffman codes are not difficult to calculate, they may represent more of an investment in math than we want to extend at this point. A simple example of the value may be obtained by looking at an empirically developed variable length code – Morse code:

._	a	_ _.	g	_ _	m	...	s	_._ _	y
_...	b	....	h	_.	n	_	t	_ _..	z
_._.	c	..	i	_ _ _	o	.._	u		
_..	d	._ _ _	j	._ _.	p	..._	v		
.	e	_._	k	_ _._	q	._ _	w		
.._.	f	._..	l	._.	r	_.._	x		

Note that the letters that occur most frequently (e and t) have the shortest codes (. and _). Similarly, the letters that occur least frequently (q and z) have the longest codes. Shannon's work has led to a great amount of theoretical work on information – optimal code sizes, perfect encryption, etc.

Other Measures of Information

Shannon conducted other experiments, both thought and formal, that help us begin to get a measure of information. A personal favorite is a very simple one that goes like this. First, understand that the probability of letters in English described above is not the total story. While it is true so far as it goes, the more interesting story about language is the redundancy built into it. Given a letter in English, the probability of the next letter is not the probability of the letter reported above. So, if I tell you the letter "t" appears, most of you will agree that the probability is that the next letter will be an 'h" rather than a "q." We could calculate digram, and trigram frequencies to provide for more efficient codes. But Shannon's simple experiment is still more provocative. He provided a part of a sentence, such as:

"*one can relate it to the kinds of mes_*"

or

"*if a communication channel can carr_*"

What he found is that the probability of predicting the next letter correctly was about .5. I suspect that you had no problem seeing that the next letter in first fragment is likely an "s" and in the second fragment it is a "y." Obviously, other fragments might be less certain leading to some errors. This means, think about it, that the amount of information in each letter in the sequence is one bit. Now that suggests that rather than needing 4.7 bits per symbol or 4.1 bits, in a message consisting of natural language, the ultimate compression of the message might be as little as 1 bit/letter! (Interestingly, this observation is used in some compression techniques where a clever method is used to "predict" the next symbol and what is actually encoded is the number of times the prediction is correct in a row or incorrect in a row.)

Another example of the redundant nature of natural language is demonstrated in the following paragraph. Can you read what it says? All the letters have been jumbled (mixed). Only the first and last letter of each word is in the right place:

> *I cnduo't bvleiee taht I culod aulaclty uesdtannrd waht I was rdnaieg. Unisg the icndeblire pweor of the hmuan mnid, aocdcrnig to rseecrah at Cmabrigde Uinervtisy, it dseno't mttaer in waht oderr the lterets in a wrod are, the olny irpoamtnt tihng is taht the frsit and lsat ltteer be in the rhgit pclae. The rset can be a taotl mses and you can sitll raed it whoutit a pboerlm. Tihs is bucseae the huamn mnid deos not raed ervey ltteer by istlef, but the wrod as a wlohe. Aaznmig, huh? Yaeh and I awlyas tghhuot slelinpg was ipmorantt! See if yuor fdreins can raed tihs too.*

A final example is taken from Bob Lucky's book **Silicon Dreams**. He provides a more detailed example, but the basic idea is simple. In a natural language, as we discussed above, letters appear with different probabilities. Further, once we choose a letter, the probability of the next letter is impacted. For example, once I choose a 'q' randomly, it is much more likely that the next letter will be a 'u' than a 'z'. So the second row – first order, chooses each letter based on the probability of the second letter following the letter before it. And the process goes on. Given two letters, "th", we can look to collections of writing as assess the probabilities of what the next letter will be. After four letters, there may only be one or two letters that can occur. Keep in mind, we pick all the letters at random, but based on the probability the letter will occur. A simple way of thinking about this is as follows. If the probability a 'q' will be followed by a 'u' is very high and the probability of a 'b' is very low, we might put 90 'u's in a bag and only one 'b'. Then we will pick a letter from the bag randomly. The first, zeroth order line has one letter of each type in the bag from which one is selected randomly and then put back. On the fourth order, we place letters in the bag based on the likely hood of the letters occurring after the preceding three letters.

Zeroth Order: *sfptroy efhubxiebbb ghiocwfbsgi fafqqalf*

First Order: *yos smyhota I n ssbniletns anoosnrnop kwe*

Second Order: *e obutant snwe o mar tbionas pr is withious*

Third Order: *thea se thook somly let sher of mory*

Fourth Order: *job providual better trand she displayed code*

The last line is not Shakespeare, but there is a clear difference between the random letters in line one and those where the selection of a letter is based on the three letters that precede it. This demonstration supports Shannon's observation that in context, there is very little information in each of the succeeding letters in a natural language sequence.

Information as Commodity

If I can get information, or if I have some information you might want, it is hard to avoid thinking of information as a commodity. Information is no different than physical objects that are differentially valued by individuals. However, the more interesting aspects of information as a commodity relate to how it varies from physical commodities. Consider just a few of the differences.

For physical commodities singular ownership is the norm. If I own a car and loan it to you, then you now have the car and I do not. But if I give you some information, it does not deny me the information – we can both "own" the information. Closely related to this is the margin cost of duplication. That is, ignoring the cost of creating a physical or information commodity, there is a marginal cost associate with duplication of the commodity. For information, the marginal cost of duplication approaches zero. Further, if I use information, or share it, the impact of that use on the value of the information is unclear. In some cases, sharing information may increase the value of the information. The more people who have knowledge of the facts related to climate change, the more valuable that information becomes. On the other hand, the more people who have information about what horse will win a race, the less valuable that information becomes – at least in terms of the potential return on a bet.

Information and Noise

The definition of information also allows for a definition of noise. Noise is defined as random signals in a communications channel. Most importantly, the capacity of a channel to carry information is limited by the amount of noise in the channel. If a channel is "noisy" steps must be taken to add redundancy to the message to insure it is not misinterpreted at the receiver. Simple examples are easy to create. Consider a message as a series of letters. If I send the message "help" and the channel replaces the "p" with a "d" the receiver gets what appears to be a valid message, but it is not. Now imagine, as actually occurred early in telecommunications history, I send "hheellpp" for the message help. If I get the message "hheellpd", I would know there is an error, but I would not know which word had actually be sent. If I sent the message "hhheeelllppp" and got the message "hhheeelllpdp" I could conclude under certain assumptions about the nature of noise that the actual message had been "help." Error correction techniques are quite more sophisticated than redundant transmission, but in general they are all based on assumptions about the nature of noise that have been empirically validated.

This introduction to information and noise is only the beginning, and while noise is a very important concept in telecommunications, it becomes an important concept in far more general uses. Consider the case of cross selling on Amazon.com. How does that occur? Why when you buy one product, does Amazon suggest another for your consideration? If Amazon were to make stupid suggestions, at best their sales would not increase. At worst, they might frustrate you and drive you away from their site. What is Amazon doing? Well, they are looking at millions of individuals and their purchasing patterns. Some of those patterns represent information, and others represent noise. The goal of the programmers is to separate the information from the noise. The techniques they use classify "random" patterns from high probability patterns.

The Process of Informing

Most of the discussion to this point has been about information as a thing. We have also alluded to the fact that there is a process by which information is created and consolidated. In pure form, an individual can make an observation about the world that brings new information to them. The golden age of deriving information from data aggregated from signals obtained from the real world has long passed, although there are surely still some examples – such as the discover of distant stars with planets based on the wobble observed in the brightness of the remote star as a planet moved in front of it. I would think that some of the examples of the golden age might include observations such as determining the curvature of the earth based on line of site observations. The moment when someone observed that a ship disappeared over the horizon and used that data to calculate the curvature of the earth and then the diameter of the earth must have been heady moments. Noting that there were "stars" that moved through the sky against a field of "stars" that didn't move and realizing that these moving stars were actually planets is a similarly impressive discovery. One could include other examples such as the displacement of water, or the impact of gravity on falling objects.

In today's world this kind of observation accounts for very little of the new information that is produced. The most brilliant contributions occur in the extension of what we know into new frameworks. That is, rather than information generated from observing the world in raw form, new information is generated from our existing store. More often today, new information come from theoretical manipulations such as Einstein's thought experiments related to individual frames of reference and space-time. Keep in mind that Einstein's theoretical extensions were several years in advance of experimental data that confirmed the validity of the postulation. Einstein's thought experiments are rivaled by Shannon's related to information. The insight provided by the formula presented above represents a brilliant observation about our world of information.

Shoshanna Zuboff, in her brilliant book **In the Age of the Smart Machine**, suggested that we might use the term information as a parallel to the term automation. A professor of business at Harvard, Zuboff suggested that just as power technology was used to automate work processes, information technology might be used to produce information streams that "informate" the work process. Another contributor to this notion of information as activity was the late Michael Dertouzos, former director of the MIT Laboratory for Computer Science. In his book, **What Will Be: How the New World of Information Will Change Our Lives**, Dertouzos suggested two concepts which are important to this discussion. The first is the notion of electronic proximity. Put simply, the idea is that the world is more tightly knit by the evolution of information. The second concept he puts forward is the idea of "electronic bulldozers." Basically, he observes that the tools currently available to us do little to empower us. He suggests that the electronic tools that have been developed to date are more like shovels than bulldozers. That is, our computers help us, but they are still primitive and crude and require more work than they should. These two ideas together suggest a new process to which information is being applied.

Today, we observe the development of more and more powerful tools that make use of encoded information. Consider the spell checking software in word processors. It used to be that we would run our text through a separate spell checker – indeed it was not long ago when editing, spell checking, grammar checking and formatting were all separate processes. Today, the difference, in the example of spell checking, is the amount of information that is being brought to bear. We can encode rules that distinguish between too dollars, two dollars, and to dollars. Indeed in a good spell checker the rules will correctly distinguish between the conversion of “Euros to dollars” and “the cost of two dollars.” This advance in spelling and grammar checking has been extended to speech recognition and many other digital processes. We now apply tremendous processing power, at several levels of abstraction, to these processes. We have “informated” our word processors. This process is relatively invisible to us, and some of us may miss the awesome power brought to bare. Keep in mind, these multiple levels of information and knowledge applied to information processing are emerging all around us. Every time your credit card account is charged, a variety of filters – information enriched processes – are applied to see if the charge is valid. Banks make loans based on rules applied to the application, student registrations are processed based on informated processes, and tax returns are audited when thresholds are exceeded. We will address some of the goals and visions of this process of information in later chapters. For now, it is sufficient to suggest that this may well be looked back upon as the golden age of informating digital processes. Most of the gains we are making at this time are noticeable and significant. In the future, the changes will be less dramatic and will build incrementally on what we have achieved over the last two decades. Most importantly, while we will find some algorithms that are superior to those we are discovering today, the process is essentially one way. Digital tools will only get better – there is no going backwards.

Chapter IV: Digital Information

Introduction

This chapter provides a very brief a primer on digital information and what makes it different. One of the mantras of the digital information revolution is the significant advantages of bits over atoms. While bits will not weave a warm sweater or shield us from the rain, when it comes to information they have some significant advantages over the representation of information in atoms. There are also some potential limitations related to the processing of information in digital form. This section reviews these advantages and limitations.

Key Advantages of Digital Information

The first advantage of digital information is the speed with which it can be manipulated. This speed is both linear and more recently massively parallel in nature. As a matter of fact, both speed and manipulability can be separated out as advantages. It is important to also keep in mind that not only can we manipulate digital information more quickly, the kinds of manipulation that we can do are almost without limitation. Consider just a few simple examples.

Once text is in digital form, we can find all instances of a given word in just a second or so. Indeed, as I am typing this text, I am given a real time snapshot of the number of words, and the number of pages. With a single click I can learn that as of today I have written 85 pages containing 29,401 words, 869 paragraphs, and 2785 lines. If I wanted to change the formatting of the document, I could do either the entire document or just a portion of it in a second or two.

As another example, consider a digital photograph consisting of several million pixels. A recent vacation photo with a resolution of 4000x3000 pixels containing color information of 24 bits per pixel opens in just a second. That’s 36,000,000 bytes of information! I can reverse every color – create a negative, resize the image, take the red-eye color out of a portion of the image, crop a section or undertake any other operation that might be desired in just a second.

Error Correction

As we discussed in the last chapter, the quality of information is impacted by the amount of noise introduced into the communication channel. This noise exists in both the analog and the digital world. Dust on a photograph, water damage to a handwritten communication, and paper deterioration over time are all examples of noise in the physical world. Those of us born in the era of analog phones remember the noise present in a long distance call, the snow on a television screen, or the reflection of television signals off big buildings. We have alluded to the fact that Shannon's work led to techniques that allowed us to detect and correct errors. Error correction techniques to preserve digital information are numerous, and while the subject is the matter of long and complicated mathematics, we can provide a simple sense of how it works using a simple example.

In Chapter 3, we talked about using redundancy to overcome the noise in a communication channel. Error detection and error correction make use of the fact that noise tends to be isolated sporadic events. That is, experimental observation tells us that noise consists of infrequent and isolated events. If we were to find that events occurred in closely linked batches, we would revise the techniques to overcome them. For now, we demonstrate the simplest form of error detection and a simple mechanism for detection and correction. One of the earliest error detection techniques involved what was called a parity bit. It worked for both the storage and transmission of information. Consider a message to be sent consisting of the bits:

0110110

If I was using even parity of error detection, I would count the number of bits that were one's and add one more bit to make the number of one bits even. Thus, at the end of my message, I would add one more bit – in this case a zero to make an even number of 1 bits. The transmitted message would be:

0110110+0

If there had been an odd number of "1" bits, the last bit would have been set to 1 to make an even number of 1 bits. Thus, the following would have been the result:

The message 1001100 would become 1001100+1

At the other end of the transmission, or when storage is read, the number of 1 bits would be read and an error flag would be raised if the count was not even. The reader who observes that this method does not reveal the location of the error nor would it work if there were two or more error is absolutely correct. The point is that a very simple method that can be understood by a grade school student can be devised to detect errors. Before reading on, consider if you can devise a scheme for not only detecting, but correcting an error.

Linear codes, such as the parity bit code, are more generally called Hamming codes, and have the ability to detect and/or correct various levels of errors. While Hamming codes are not hard to understand, it takes a little bit of patience, and we will therefore show a simpler block code which is a little more intuitive. Again assuming only one error, we will take seven messages and put them in a block with a parity bit for each row and each column. The message transmitted is on the left. The message received is on the right with the location of the two parity errors shown. The intersection of the two errors indicates where the correction is to be made.

1	0	0	1	1	1	1	1
0	0	0	0	0	1	0	1
1	1	1	1	0	0	1	1
1	1	0	0	0	0	0	0
0	0	0	0	0	1	0	1
0	1	1	1	1	0	0	0
1	1	1	1	0	0	0	0
0	0	1	0	0	1	0	0

1	0	0	1	1	1	1	1
0	0	0	0	0	1	0	1
1	1	1	1	1	0	1	1
1	1	0	0	0	0	0	0
0	0	0	0	0	1	0	1
0	1	1	1	1	0	0	0
1	1	1	1	0	0	0	0
0	0	1	0	0	1	0	0

Fixing errors introduced in the digitization of information, or in information converted from analog sources can also be astoundingly easy. Consider the following image of my son at a local park. It was scanned from a negative that had been handled and had a finger print on it. Figure 9 shows the photo, a close-up of the fingerprint over the cotton candy, and a super close-up of the dirt or noise that constitutes the fingerprint. I applied a filter to the image area to remove the noise in the image. That is shown as the image fragment in the lower left corner. Note that I have not taken out all the noise, just reduced it to a level that is less disruptive to the human eye. How was it done? The answer is amazingly simple and again depends on understanding something about the real world. Basically, when we sample an image at high resolution, the samples created by optical lenses are such that transitions are gradual.

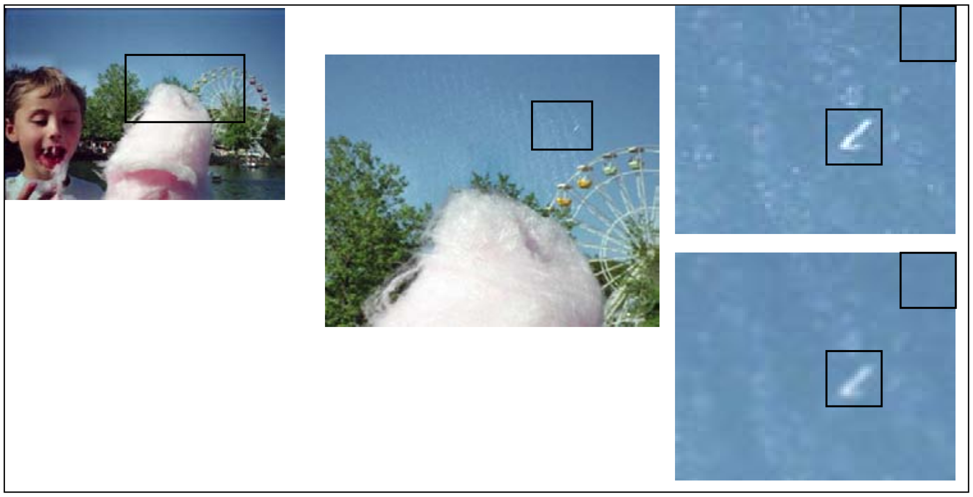


Figure 9: Removing Noise in an Image

Figure 10 shows two close-ups of the image. The first is 388 x 313 pixels. The second is 51 x 28 pixels. What you can see, even though our software is working against us to smooth the image, is that what appears to be an abrupt transition in the original image is actually a relatively smooth transition from the blue of the sky to the grey-green of the Ferris wheel.

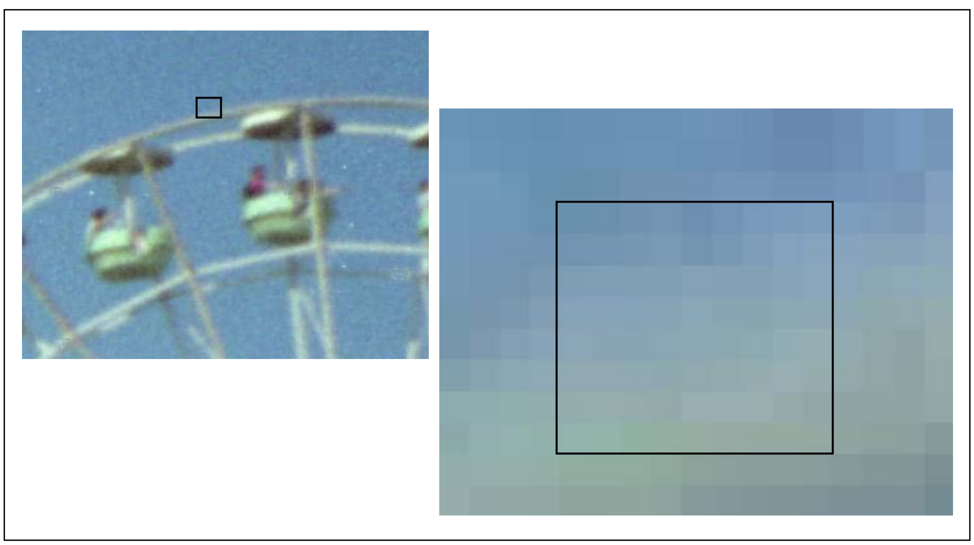


Figure 10: The reality of Sharp Edges in Images

If we were to examine the numeric values that are stored in the file, what we would find would be a series of numbers that change relatively slowly. While the pixels in these color images are actually represented as triplets of numbers –representing the red, green, and blue values, we use a single number – as if the image were made up of shades of grey. The 8x8 pixel area where the color transitions from the sky to the Ferris wheel above contains numbers like those shown in the table below. Note that the numbers over a small area change very gradually.

100	100	100	100	100	100	104	104
100	100	102	102	102	104	105	105
102	103	104	104	105	105	106	106
103	105	105	106	106	106	107	107
105	107	106	107	107	108	108	108
106	107	107	108	108	109	109	109
107	108	108	109	109	110	110	110
108	109	109	110	110	111	111	111

Now look at a second table, shown below. It is identical to the one above with the exception of two cells – which are highlighted in yellow to make them easier to find. In a real image, with very rare exceptions, these pixels would represent some form of noise. Why? Because we can safely make the assumption based on real world experiments that the numeric transitions between pixels are smooth in an image captured from the real world. We can use this fact to our advantage. In the case of noise, we can apply a “filter” that takes the average of the values of all of the pixels around a given pixel and substitutes that value for the existing value. In the case of the pixel with a value of 216, the calculated value would be 106.625 which would round to 107, the original value. There are simple filters to smooth images, and sharpen them, and detect edges, etc.

100	100	100	100	100	100	104	104
100	100	102	102	102	104	105	105
102	103	104	104	105	105	106	106
103	105	105	106	106	106	107	107
105	107	106	216	107	108	108	108
106	107	107	108	108	109	109	109
107	108	108	109	109	110	215	110
108	109	109	110	110	111	111	111

As you might guess, we can remove noise in almost anything represented digitally. Indeed while there have been many improvements in phone technology that eliminate the noise in the channels, it is the conversion to digital transmission that is responsible for the majority of the noise reduction that allows you to hear a pin drop.

One final point about error correction. Correcting errors takes time and compute cycles. It can be more of a disruption – caused by a processing delay – than just letting the error go by. This is indeed what is done with errors when reading an audio CD. If a scratch on the CD causes an error, the processing software does the kind of averaging discussed above to smooth over the sound. At other times, errors are unacceptable. In this case we need to have really good error correction. Data CDs must be able to correct the errors in the data stored on the CD. For this reason, about 15% of the bytes on a data CD are devoted to error detection and correction – they use a Cross Interleaved Reed Solomon Code. Standards for the longevity of data recorded on CD takes account of the ability of the error correction data to account for noise that occurs over time.

Encryption

Most of us have played games with encrypting messages. In some ways, very complicated encryption is nothing more than complex variations of two simple and easy to understand encryption methods – substitution and permutation. Encryption by substitution is an old technique and one that many school boys of my age practiced using a secret decoder ring – be it the Captain Midnight, Sky King, or Jonny Quest variation (shown) where the ring had a disk that could be rotated to identify the cipher text to be substituted for the plain text message. Some substituted numbers for letters, others used letter variations, but in both cases, the substitution was based on the simplest possible algorithm – a numeric offset ('a' was replaced by 'b', 'b' by 'c', etc.). The message:

"hello how are you"

offset by 3 with becomes:

"khoor krz duh brx"

If we include spaces as the letter just past z, the message becomes:

"khoorakrzaduhabrx"

Some readers may see the weakness of this technique based on some things we have already discussed about language. Specifically, if we know that the most frequently occurring letter in our cipher text is "q" it is a pretty safe bet that if we substitute 'e' for 'q' and continue based on letter frequency, we should be able to decode the message. Indeed if the cipher is sufficiently long, algorithmic decoding will normally be between 75% and 90% correct based simply on the probabilities.

This leads us to seek out another technique for encryption – permutation. Let's not bother to substitute, but simply jumble the message. Again we offer an exceptionally simple example of a permutation. We begin with the same "hello …" message, and write it horizontally into a five by five table:

h	e	l	L	o
	h	o	w	
a	r	e		y
o	u			

and then read it out vertically, giving us the ciphertext:

“h aoehrulw o y   ”

We can vary the size and shape of the block or create some other permutation to hide the message.

Sometimes we define only these two basic techniques, other times we include a key as a third technique. The reason is that the use of a key simply provides a variable kind of substitution. Consider, again over simplistically, that we modify our substitution algorithm in the following way. For the first letter in our message, we will substitute a letter three positions away, for the second letter, we will substitute a letter 1 position away, for the third letter, we will substitute a letter five positions way. For the fourth fifth and sixth letter, we will repeat the process. In cryptography, we refer to “315” as the cryptographic key. We can no longer use simple letter frequencies to break the substitution code, because “e” will sometimes become ‘h’, other times ‘f’ and yet other times ‘j’. As the key becomes longer, the ability to use letter frequencies as a technique decreases. It is still possible to break the cipher, but it becomes increasingly difficult. Modern encryption techniques use a combination of substitution and permutation based on the use of a sufficiently long key to make it “reasonably difficult” to break the code.

The discussion of encryption will be important in later chapters when we talk about privacy, authentication, non-repudiation and other important aspects of having information and doing business in the digital environment. For this simple technology introduction, it is useful to introduce single versus dual key encryption. If we return to our everyday understanding of encryption, most of us will be able to dredge up some movie experience related to encrypted messages, or maybe you know the story of breaking the code for the German “enigma machine.” Settings on the machine constituted a cryptographic key. These settings were changed each day based on information contained in a codebook which was carefully guarded. If the daily keys were captured, and messages transmitted using them could be easily decoded. Whether used for diplomatic messages or military communication, the most important part of the process was insuring that no one except the sender and the receiver had the secret key. If the messages are between preselected parties – the generals, or the president and the ambassador, we can manage to arrange to have the keys secured by the parties. However, if we need more widespread exchange of information – say between me and a bank, we need a more ad hoc method of exchanging keys. The solution to this problem is to use a method called dual key encryption. Again, we will discuss it in more detail in a later chapter, but consider the simplest scenario, exchanging a message between you and me. Each of us will build a key pair. The nature of the construction process is complicated, but basically it results in the ability to use either of two keys to encrypt or decrypt a message. One of the keys is made public, the other kept very private. I prepare a message from me to you and encrypt it using my private key. I take the result and encrypt it using your public key. Now I send the message to you. If anyone else intercepts it, they can do nothing with it unless they have your private key – which they don't. Having your public key does them no good. So you get it and decrypt it. But what you get out is still encrypted. You use my public key to decrypt it. This gives you an assurance that the message is from me. (Someone else could have sent you an encrypted message, but my public key will only decrypt a message from me.

One final note on this brief introduction to a rich and complex topic. You will recollect that Shannon researched the relationship between information and noise in a communications channel. If a message is corrupted by "noise" it reduces the capacity of the channel to carry information. When we apply a set of substitutions based on a key to a message, we introduce noise or randomness. Shannon proved that a key of length equal to the message that was guaranteed to be random would be unable to be broken. The problem here is that it is nearly impossible to produce a truly random key! But if we can produce one, we are assured of being able to encrypt a message that can only be decrypted by the same random key.

Compression

Compression techniques may be more interesting than either error correction or encryption. It is difficult to communicate the importance of compression to most students of information because of the vast strides made in increasing the bandwidth of our networks and the storage capacity of our devices. At the same time, compression is always important at the cutting edge – i.e. today compression is important to increasing the speed of cellular systems and digital video. To get a sense of how compression works, consider how it impacts fax and digital photos.

You have a fax machine in your office or home that transmits images over phone lines. For a variety of reasons, standard phone lines can transmit no more than about 35,000 bits per second, and the standard for fax machine transmission is set at 9,600 bps. (This limit does not exist for ISDN[13] or Digital Subscriber Lines (DSL) or more modern cellular phone connections, but the least common denominator for fax worldwide is a

[13] ISDN stands for Integrated Services Digital Network, an early standard for substituting digital service across the public telephone network.

standard analog line. Let's imagine that you want to transmit just one page at high fax resolution – which is about 200 dots per inch, both vertically and horizontally. That means we need to transmit 200x11 lines where each line has 200x8.5 pixels. That means we need to transmit 1700 pixels/line times 2200 lines, or coding for 3,740,000 pixels. At 9,600/second, a fax pages will require 6 minutes and 30 seconds/page. The reality is that we don't transmit the page pixel by pixel. For the widespread CCITT[14] Group 3 fax, we use variable length code – a type of Huffman code – that was developed specifically for fax transmission. It is optimized for long runs of white codes and short runs of black codes. The code is actually made up of two types of codes – makeup and terminating. There are further complexities in how the code is put together, but you can get a flavor of how it works by looking at some of the codes associated with various pixel run lengths:

Run length	Code
2 black pixels	11
3 black pixels	10
62 black pixels	000001100110
1664 white pixels	011000
64 white pixels	11011
26 black pixels	000011001010

If we imagine normal pages made up of long runs of white at the top, bottom, and sides of a standard business letter, we can see how this code would significantly reduce the number of bits that need to be sent. At the extreme, using a code such as "011000" to represent 1664 bits represents a compression of 277 to 1. That reduces a 6 minute and 30 second transmission time to about one and a half seconds. You now understand why the transmission of a letter by fax buzzes through the first inch or so, slows down over the body of the letter, and then races along the blank space at the bottom of the letter. Please keep in mind that the transmission code is somewhat more complex than I have described, but the essence of the idea doesn't change. Also, today's more common CCITT Group 4 fax uses a very different compression methodology based on the fact that digitized images change slowly – remember the discussion of noise in images. A variation of the rule that images change slowly allows only the changes to be transmitted, greatly reducing the length of the message.

Today, the cost of a terabyte of storage is about $100 and it will continue to decrease. We can now buy an SD storage card for our cameras at a cost of $1-5 per gigabyte, depending on the quality of the card, i.e. the speed at which the card operates is a major factor in cost. Every time the

[14] Today, international standards for telecommunications are set by the telecommunications sector of the International Telecommunications Union – ITU-T. Prior to 1992, this sector was referred to as International Telegraph and Telephone Consultative Committee, or in French "Comité Consultatif International Téléphonique et Télégraphique" – or CCITT.

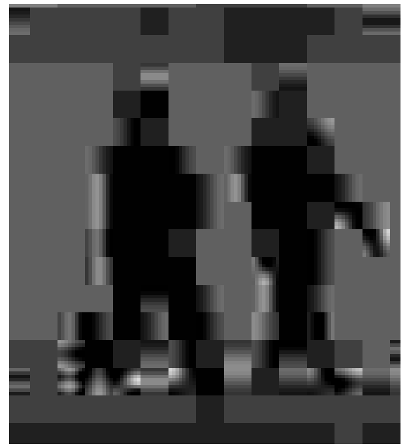
cost of storage decreases, we seem to find ways to fill it more quickly. Contemporary cameras are recording images of about 12.1 million pixels – 4000 x 3000 pixels. This provides a reasonable resolution, but can quickly fill up an SD card. For example, an uncompressed twelve mega pixel image requires 36 million bytes of storage. Using lossless compression methods – methods that allow us to recover the values of every pixel, we can reduce an image by 30% to 50% -- depending on the complexity of the image. Using lossy compression methods, where we lose some of the details, we can compress the image 50% to 90% or more depending on both the image and the selected degree of compression. To give you a sense of what this means for actual images, take a look Figure 11 which contains four copies of a picture of the backend of the Monument to the Tomb of the Unknown Soldier at Arlington National Cemetery. I chose this particular photo because the level of detail creates a difficult compression problem. On the left is an uncompressed copy(35MB), next is a lossless compression (23MB), third is a high quality lossy compression (7.8MB) and finally a medium quality lossy compression(1.5 MB)

Figure 11: Image Compression

Odds are, there is more distortion in the quality of these photos related to the printing process than there is based on the actual quality of the image in terms of the number of pixels actually being displayed. Figure 12 shows the impact of lossy compression. The four images shown below are blown ups of a very small section of the photo – a little piece that shows two people walking by the monument in the photo above – they are just to the left of the large tree in the right of the photo. Below we show a small section of the photos magnified much more than they would be in a normal printout.

Figure 12: Impact of Image Compression Using Discrete Cosine Transform

On the top left is the fragment from the uncompressed copy. Top right shows a high quality lossy compression. (We don't show the lossless compression, because it would be the same as the uncompressed image.) These two images are not identical, but even at this magnification it is hard to see the difference. The bottom two images show lossless compression at medium quality and low quality. You can see the result of the Joint Photographic Expert Group (JPEG) compression algorithm. If you look very closely, you can see sets of pixels (actually 8x8 squares) that are visible. At the extreme, and as an oversimplification of the algorithm, these 8 by 8 groups of pixels have little or no internal variation. Keep in mind, the uncompressed image was 35,000, 000 bytes. The high quality compression photo was 9,000,000 bytes, the medium quality was 1,500,000 bytes, and the low quality image was compressed to 850,000 bytes. That means the highly compressed photo, which may be identical to the casual gaze, requires less than 2.5% of the storage space of the uncompressed image. That means you can store more than 1200 high resolution images in the same space as 28 uncompressed images! For those of you who are



curious, this image, under maximum JPEG compression produces a small (200,000 bytes) but acceptable photo for most uses, and results in our two visitors looking as shown here.

As with cryptography, compression is an area of study in its own right. It easily occupies a full course in a graduate program, and personally I find it one of the most enjoyable areas of study in which I have engaged. The enjoyment comes from the fact that good compression techniques often reflect the meeting of mathematical processing with insightful observations about the real world. A good example of this is the marriage of solid information theory about variable length coding and the observation that the most common form of fax transmission include long runs of white and short runs of black. The variable length code was used because there is a significant variability in the probability of certain patterns of white and black pixels on a number of sample letters that might be faxed.

One final note about compression. Some compression techniques are married to a particular form of digital information, other are more general. Some forms of compression are “lossless”, others are “lossy.” That is, we can sometimes achieve dramatic increases in compression by representing it in some specialized form that does not capture every aspect of the original. We will discuss a little more about this aspect of digital information when we talk about some of the limitations below.

Some Limitations of Digital Information

While the advantages of digital information are significant, there are some limitations as well. The lack of infrastructure stability has been the greatest concern to date. Moving forward, decisions about fidelity will be the greatest concern.

Infrastructure Stability

For the last 2000 years, messages placed on paper like substrates with some form of ink have been the norm for communication. There are no doubt issues with crumbling papyrus and ink that fades, but to a large degree, those messages produced through this period have been stable. The same cannot be said for documents saved to magnetic tape or floppy disks. Over the last thirty years, I have actively migrated documents across a series of infrastructures – from 8 inch floppies and magnetic tape to 5.25 and then 3.5 inch floppies. Now, they are all stored on hard drives and

CD's. I still have the 8 inch floppies that were used in the Xerox Star in the early 1980's. Even if I still had a drive, the formats used by Xerox have long since fallen into disuse. Over the last thirty years, there was one standard format for characters – the American Standard Code for Information Interchange (ASCII). Even if I have to deal with the fact that the embedded word processing codes were no longer meaningful without the word processor they were meant for, I can still create the basic text message. (ASCII is still used as a part of a new standard for character coding called Unicode, which makes it possible to represent the characters of almost all of the world's languages.)

We are still searching for a millennium format that will survive the ravages of time. With XML and Unicode, we may have such a base. At the current time, the hegemony of Microsoft presents a real concern. The quality of Microsoft software coupled with the relative stability of the company hides the fact that the formats used by Microsoft are binary and proprietary. Unless I take action to save my word document or excel spreadsheet in a more stable and standard form, that information can be lost without the software to interpret it. Given the ubiquity of Microsoft products, it is unlikely that users will be stranded, but it is by no means impossible. Even if we can save the basic information, it is unlikely that things like a power point animation will be saved to more standard formats. Indeed many of the things that make Microsoft so special are most subject to easy loss in migration

We also need to be concerned with migrating information from one media to another. We are getting better at understanding the need to migrate files, but it is still not widely viewed as an individual or organizational responsibility. Given the increases in processing power and the decreases in the cost of reliable storage, we are much further along today than we were a decade ago. At the same time, having migrated all my super 8 film to VHS tapes and having migrated all my VHS tapes to DVD's, I will still need to decide when I need to migrate the DVD's to BluRay or the next generation of storage. It takes time and costs money.

Decisions about fidelity

A more subtle issue for digital information relates to the fidelity of the information. The obvious question has to do with the number of samples that are taken. An image at 300 dpi provides a reasonable level of resolution for most uses. Indeed given the optics in most cameras, any increase in resolution might be wasted. Early digital cameras had resolutions at the one megapixel level. Once these images are captured, there is no way to capture the information lost by the averaging of information over the area captured by the sensor. You can interpolate, but the data is being inferred not captured. So the first question about fidelity is: when is the digital fidelity sufficient to replace the analog sample? Some people still argue that there are nuances captured by a vinyl record, which is analog, versus the digital representation obtained in a CD.

The general rule of thumb is set by measurements of the human sensory system. For example, humans are capable, on average, of hearing sounds in the range of 20-20,000 Hz. Therefore, when we record sounds, based on the Nyquist- Shannon sampling theorem, we capture 44,000 samples per second. Basically, the Nyquist-Shannon theorem suggested that we have to sample signals at twice the rate we wish to reproduce. However, if you are the rare human that can hear sounds in the range that dogs hear them, you will not hear those sounds coming from a CD – they will be lost. For vision, at a standard distance, we see a continuous image when the number of pixels per inch exceeds 400. In reality, that number is lower when we are looking at 1000's of colors. That resolution, 400dpi, is a good standard for pure black and white images. This basically suggests that an 8x10 photograph composed of 3000 x 4000 pixels is well above this fidelity level. Again, not to belabor the point, but if we wish to be able to determine the color of the eyes of the two people captured in the photo of the Tomb of the Unknown, we will need a lot more resolution. When the use is archival, there can be extended discussion of how much resolution is acceptable, particularly if the only copy saved is digital.

One last point about fidelity. Fidelity is not only dependent on the number of samples; it is dependent upon what each of those samples represents. Consider these two images of my younger son playing soccer shown in Figure 13. For most of us, they look identical. As a matter of fact, the compression method used on the image on the right reduces the number of unique colors from 115,800 to 256 in the image on the left.

Figure 13: Image with True Color and Reduced Color Palette

This is made clearer in the blow up below of my son's face. Normal viewing of most images shows very little, especially when the range of colors is relatively limited – such as in a portrait. However, when the sampling algorithm is forced to choose the most dominant colors from a large palette, the impact can be significant. Of some curiosity is the fact that the compression algorithm used on the left is considered "lossy" while the one used on the right is considered "lossless." The fact is that this is correct related to the compression, but not accurate in terms of the fidelity of the sampled image.

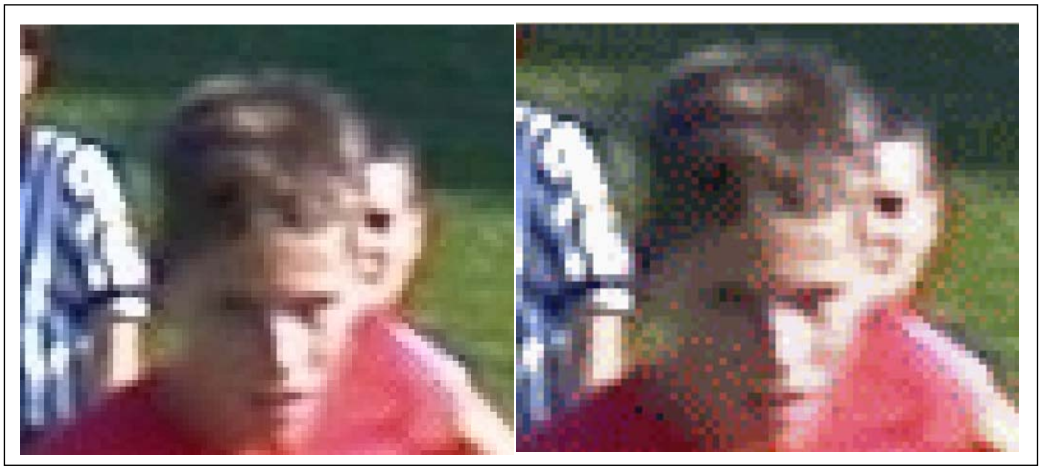


Figure 14: Impact of a Reduced Color Palette

There is a similar, although more subtle, issue related to symbolic information. Many of us have typed foreign words that required the use of special characters as their Latin alphabet equivalent rather than dealing with the issues associated with getting the correct symbol in the document. Using a simplified symbol set, we have lost fidelity.

Chapter V: Technology Simplified

Introduction

This chapter is broken into two parts. The first part looks at technology trends with an eye to predicting what new developments will be important. The second part looks at the technology that underpins the World Wide Web, again with an eye to explaining both what has happened and where it is headed.

There is little doubt that we are in a period of rapid change. What is sometimes less clear is that it implies significant experimentation. As we discussed earlier sometimes new technologies are thought about in terms of what they replace. The automobile was originally a “horseless carriage” and radio was “wireless telegraphy.” We are still working with cordless tools and wireless networks. We are still concerned about, or doubtful about the paperless office. All of these and other less obvious ways of thinking limit how we imagine the use of new technologies. Indeed, I will explore in the next chapter how we might re-imagine documents once we free ourselves of thinking about them as paper-based entities

There are many ways to expand on the implications of such an attitude, but I have a personal favorite that I hope will be illuminating in the rich and complex fabric of the story. Johannes Gutenberg was a goldsmith born around the turn of the fifteenth century. In his late 40’s, after a number of experiments in printing, Gutenberg finally managed to perfect a method for moveable type printing. With loans from his brother in law, he was able to get started. In 1450, he borrowed some money from Johan Fust, and began to work with a scribe, Peter Schöffer, who is thought to have been the type designer. The project they set out on was the

printing of about 200 copies of a bible – now referred to as the Gutenberg Bible. Note that each line of the Gutenberg bible shown on the previous page is justified and that the page is embellished with graphics. This was an amazing accomplishment. Programs to justify text are very difficult to write – doing that kind of variable spacing in metal type is very difficult – as is the inclusion of graphics. A large part of the value of these marvelous documents is the technical sophistication of the accomplishment. Why make this task of printing words so difficult? Because he wanted to create a mass produced document that would be considered equivalent to a handmade document – a bible copied by a monk. Why did the monks justify text? Because the copying was an act of prayer and justified lines was considered to be a more perfect tribute. (The fact that justified text is harder to read was not a consideration.) Similarly Gutenberg and Schöffer produced typefaces that mimicked the hand produced letters of the monks. It was about 50 years later that Aldus Manutius introduced a lighter italic type that made smaller lighter books possible.

The point of this story is that resistance to new technologies may cause us to take extraordinary steps to think of them or sell them in the context of old ways of doing things. It may take a generation or more for people to be freed from thinking of the new technologies in the context of historical processes. Indeed Aldus Manutius was born into the world of mass production printing – 1449. He was a child of the new technology.

It is interesting to note that students of this period refer to the books, pamphlets and broadsides produced between 1450 and 1500 as incunabula. The term is Latin for "swaddling clothes" or "cradle." Webster's New World Dictionary defines it as: “The very first stages of anything; infancy; beginnings.” More broadly still, incunabula are the products of any art or industry in the early stage of development. I purchased my first portable computer somewhere around 1984 – I forget the exact serial number which was hand printed on the circuit board, but it was less than 200. The Compaq portable was a good example of incunabula. It weighed about 30 pounds and was pretty much shaped like a sewing machine. It still required a power outlet, but you could carry your computer with you. Incunabula are the products of a period of significant experimentation. During these early days of experimentation, there are few standards, conventions, or traditions. The innovators are free to mold the new products after existing objects, or they

can take radical approaches to design – designs that in the light of eventual designs seem very strange. For example, we are still in a period of significant experimentation with cell phone technology in terms of form and function.

Selected Definitions

I tried to write this chapter with a minimum of technical definitions, but in the end I was not entirely successful. After looking at this chapter and the next, it became apparent that I needed to use some technical terms that make things easier. You can work on these now, or come back to them later as you see a need. The list is very selective and designed only to serve the needs of this book.

Protocol — A network of connected computers is bound by protocols. They are everywhere and often sets of protocols are involved in every transaction. A protocol is, just as in the world of diplomacy, the rules that govern an interaction. In its simplest technical implementation, fans of CB radios will recall that each utterance on the radio was ended with the word "over" which signaled our partner that we were done and handing control "over" to them. When a conversation was completed, either speaker could signal that by using the phrase "over and out." While it may require hundreds of pages to describe a protocol, to insure that there is no miscommunication, the guts of a protocol define who is allowed to initiate and terminate interactions, how do they signal such, how is control exchanged, how are messages formatted, what can the various parties say, and how are problems handled. It is that simple. These "simple" protocols are defined in standards that normally run between 100 and 300 pages.

There is one final point about protocols that is important to consider. While there are what we call "peer-to-peer" protocols and "broadcast and multicast protocols", by far the most common kind of protocol model is the client server model. Under this model, one of the communicating programs is started and programmed to wait for a connection from another program. The program, or more correctly process, that waits for connections is the server, sometimes called a daemon. The program or process that initiates the communication is called the client.

Addresses — Most of us are aware that every computer in the world has an address. Some may be aware that today the public world uses an addressing scheme defined by the internet protocol.

In addition, some will be aware that the current 32 bit addressing scheme is scheduled for migration to a 128 bit addressing scheme. If you understand binary numbers, great. If not, trust me when I tell you that a 32 bit number can be used to name 4 billion (4,294,967,296) machines. Version 6 allows the identification of 340 unidecillion machines. (That would be 340 followed by 36 zeros!

Returning to addressing generally, there are some very special things about IP addresses. Let's look at three addresses in a sequence.

10001000 10001110 01101010 00011001
10001000 10001110 01101010 00011010
10001000 10001110 01101010 00011011

These three IP addresses are a little hard to read, but they do provide a way to identify a particular machine. Note the first group of 8 bits – the first two bits are 10. This means these addresses are a part of one of the 16,384 class B networks assigned by the Internet Assigned Numbers Authority (IANA). Basically, when a class B network is assigned, it gives the owner a network in which they have the right to assign addresses for a little more than 65000 machines. At the risk of saying too much, the first sixteen bits define the class B network, and the last 16 bits are used in creating addresses for the 65,000 machines owned by that organization.

We need to take two more steps. First, keeping 32 binary digits in your head is pretty much impossible. We can convert a binary digit to a decimal digit by adding up the numbers in each position. For example, adding from right to left just as we would for a base ten number, the binary number 10010101 represents 1 1 + 0 2's + 1 4 + 0 8's + 1 16 + 0 32's + 0 64's + 1 128 or 147. If we go up to the 32^{nd} position, it represents 2147483648. To avoid this, we will represent the address as four decimal numbers separated by periods. For the three addresses above, these numbers are:

136.142.116.25
136.142.116.26
136.142.116.27

You may not think these addresses are easier to remember, but looking at addresses over many years, it is not hard to remember that 136.142 is one of the class B networks assigned to the University of Pittsburgh and the 116 subnet is one of those assigned to the School of Information Sciences at Pitt. Within the School of Information Sciences, the permanent address for my laptop ends is 26. The second, and last thing to know about addresses is that there is a system that maps one or more names in human form to addresses. A decision was made back in ancient times – the 1980's – that all machines beginning with 136.142 would have the human name "pitt.edu." Those continuing with116 would be "sis.pitt.edu" and my laptop – ending 26, would be "cport" – a name I chose to identify the "Cascade Portable." A system was set up to map names to addresses. The system is called the "Domain Name Service" or DNS. DNS makes it possible for you and I to more easily remember the address of the machine that provided web services for Sears, or JCPenny, or Amazon.

Ports — If you made it through addressing, which is hard, you will find the notion or ports very easy. When you are sitting at a computer today, it is likely that there are dozens if not hundreds of processes running. Some of them will periodically connect to another machine on the internet to exchange some information. For example, it may be the case that the clock on your computer is set to check periodically with "time servers" on the network that are maintained by the government. My machines check with a machine at the National Institute of Standards and Technology (NIST) -- "time.nist.gov." The machine at NIST might be running several processes that could answer questions and if that were the case, it would have to be able to distinguish between requests for the time and those other requests. Given that virtually all modern computers have more than 64,000 input /output ports (65,536), it is the custom to assign different servers to wait and listen for connections on different ports. The time server at NIST would be set to listen on port 37. As some of you may know, the port on which web servers normally listen is port 80. Many machines run a service called the "echo service" which by convention is run on port 7, allowing us to check to see if a machine is up and running. Collectively, the first 1024 ports on a machine are reserved for specific (privileged) services which include things like web servers, mail servers, time servers, etc.

File systems — From time to time we will talk about "file paths." Most readers will be familiar with the folders and files on their machine. We see the icons in a folder that represent other folder. Documents, or files, can be placed in a folder, but generally a folder cannot be put in a file. Given a set of folders, we can say that "hello.txt" is in the folder "examples" which is in the folder "books" which is in the folder "spring." On PC's, we have different disk drives – "C", "D", etc. On Unix, all of the physical drives are invisible and are hidden behind a "file system" that is described logically as folders or directories situated in a hierarchy that begins with a root. In the case of the file "hello.txt" and the folders, we would unambiguously define that file as a path from the root (indicated by a '\') followed by the folders or directors separated by additional '\'s. Specifically:

"/spring/books/examples/hello.txt"

The notion of file paths will be introduced as part of the scheme by which we can locate any resource anywhere in the world.

Tree

File systems are well understood structures in computer architecture because they are an example of a "tree." A tree is one of several data structures that have formal representations that allow us to perform operations on them with confidence. The dinner plates in your cabinet form a stack, and stacks have many interesting properties but the most common and easily understood is that the last item put on a stack is the first one taken out. This is contrasted with a queue, in which the first item in is the first item out. You want the grocery store line to be a queue, and not a stack. In computer programming we have uses for stacks, queues, linked lists, and trees.

More formally, we refer to a tree as a directed acyclic graph. That is, from the root of the tree, there are directed links to children, and then children of children, until we arrive at a leaf node – which has no further children. What do we mean by acyclic? Put simply, there is never a directed link from a child to a node closer to the root – a path never cycles. Knowing that the graph is both directed and acyclic allows us to make a number of formal assumptions about the elements in the tree.

DOM

One particular "tree" is of great importance in how we work with documents. In the literature on XML, HTML, and SGML, there are frequent references to the "DOM" which stands for Document Object Model. AT the core, the DOM is represented by a tree that we can process computationally in a variety of ways. Every modern browser makes use of the DOM to improve how we can process displayed documents. Early DOM's (level 0, and level 1) provided programmers very little power as they were over simplified – some parts of the tree could not be accessed. More recently, DOM's at levels 2 and 3 are providing extremely powerful processing capability.

Binary I have tried to be accurate in this book without boring the reader with strange numbers. For example, why are there 65,536 ports on a computer. The answer has to do with the binary number system. All of the concepts in this book can be understood without understanding different number systems, but somehow understanding the impact of the binary number system on strange numbers helps to eliminate the mystery. If we have one bit, we can represent two states – 0 or 1. If we have two bits, we can represent four states – 00, 01, 10, or 11. We could continue this process, but it won't take long to get to the conclusion that each bit added doubles the number of states. Put another way, if you have n bits, you can represent 2^n states. Eight bits would allow for 2*2*2*2*2*2*2*2 or 256 states. On a computer, for historical reasons, the smallest convenient unit of storage is a byte – which is made up of 8 bits. If we were to have an image made up of pixels where the color of each pixel was made up of three 8 bit codes – one for red, one for green and one for blue, the number of possible colors would be 2^{24} or 16,0777,216 different colors. If we allowed two bytes to define a port number, we would be able to define 2^{16} or 65,536 ports. The next time you see a strange number in a discussion of storage capacity, processing speed, or transmission rate, keep in mind that the specific number is less important than the general size – the specific number is most likely related to the fact that the number is a power of two.

Trends in Technology

Various people have suggested long term technology trends that have proved false and are frequently cited as suggesting the futility of trying to predict where technology will go. Some of my personal favorites include:

> *"This 'telephone' has too many shortcomings to be seriously considered as a means of communication. The device is inherently of no value to us."*
>
> *A memo at Western Union, 1876.*

"The wireless music box has no imaginable commercial value. Who would pay for a message sent to no one in particular?"
Advisers to David Sarnoff related to radio, 1921.

"There is no reason anyone would want a computer in their home."
Ken Olson, Digital Equipment Corp, 1977.

While such erroneous predictions can be fun, others have made predictions that have proved more astute. Some would suggest that Nicholas Negroponte may have been the most astute observer. Negroponte's book "Being Digital" is the inspiration for the digitization trend suggested below. He made another prediction which has come to be known as the "Negroponte flip." While it is not included directly as one of the trends I discuss below, it was so accurate and precise – despite the reservations of some critics – that it is worth a mention here. Negroponte suggested that technology would convert broadcast communications to narrowcast communications and that narrow cast communications would be converted to broadcast communications. The easiest, but not the best, way to look at this prediction is to imagine that wired phones (narrowcast) will be replaced by cell phones (broadcast). Similarly, broadcast television will ultimately be delivered by wire – narrowcast. Some argue about whether satellite TV delivery is broadcast or narrow cast. If you think about it, you can argue each way. More importantly, the many implications of Negroponte's prediction are right on. Basically, what was free has become a pay for use service. And what used to be communications tied to a given point was freed of that restriction.

Many technologies have been developed over the last few decades. Some have succeeded, others have failed. Additional technologies will be developed over the coming years. It is productive to ask how the wheat might be separated from the chaff in examining these new technologies. What does seem true is that there are trends. When technology supports these trends, it seems to take root and grow. About a decade ago, I began to review new technological developments in my classes in the context of trends such as digitization. I started with four, and ultimately added two more trends that seem to stand the test of time. These six trends are discussed in this section: digitization, accommodation, standardization, objectification, integration, and distribution.

Digitization

More and more data is being kept in digital form. From word documents to MP3 files to JPEG images, the world is increasingly digital.

It is true that more and more of the information artifacts that we create are created in digital form. It is also true that more and more of the non digital objects we create are given a digital surrogate – museum paintings, music compositions, old photographs and recordings. As we have shown above, digital data is important in that error correction techniques can be embedded in the representation to allow the data to be self correcting.

Multipurpose digital signal processing (DSP) integrated circuits are currently commonly integrated into personal computers. The PC DSP provides digital/analog conversion. They replace special purpose circuitry that is currently used to provide modem, fax, telephone (PC voice mail, fax response, etc.) and sound card functions. Features that were implemented in hardware will be implemented by software, providing for easier maintenance and enhancement.

Accommodation

Computing power has been doubling every two years. Much of this increased power has been used to improve ease of use, that is, many of the compute cycles of your devices are devoted to improving the user interface. It is interesting to note that the text based word processor of 5 years ago had almost all of the capabilities of today's WYSIWYG (What You See Is What You Get) word processor. Yet the WYSIWYG word processor requires a machine that is four to eight times more powerful. Side by side, the functions of the programs seem to operate in about the same human time. Where has all the increased processing power gone? The WYSIWYG program uses all that power to make the word processing functions easier to use. Graphical user interfaces absorbed the vast increases in computing power in the decade from 1995 to 2005. Since that time, we have made vast increases in processing spoken language and most recently touch interfaces. What began as primitive voice dialing by name with speaker training has now become speaker independent voice recognition for browser searches. The complexity of these systems is understood by few users. When I use the browser on my phone to look up Big East basketball standings, the system is not simply analyzing my speech utterances, but looking at my history of searches as well as the frequently occurring searches by all the users of the internet. This constitutes a lot of processing to look up a basketball score..

Another way to understand this technology trend is to take a broader look at what has occurred over the last 50 years. In 1962, the cost of a computer was near $1,000,000 and the average professional was making on the order of $10,000. When a conflict arose between the human and the computer, the human lost — they were the less expensive cog in the wheel. Today, when all costs are considered, a professional costs an organization closer to $100,000 while a computer of the same power as the million dollar computer of 1960 costs about $2,000. The human is now the winner when there is a conflict.

The technology of accommodation involves two areas of technology development. The first is the actual interface technologies – touch screen, mice and trackballs, speech recognition digital signal processing. Each of these technological developments is intended to provide an easier, faster, more natural way for a user to input or access data stored in a computer system. Technology is quickly getting to the point where all of these operate at a level that allows the digital representation to be indistinguishable from the corresponding analog representation, given the normal acuity of the human sensory or motor system in question.

The second area of technology development is the multi-level data mining that is being applied to human activity. We now do speech recognition by trying to guess what the utterance is in the context of words that have already been recognized, in terms of known user interests based on historical activity, and guesses based on crowd behavior – i.e. if the utterance could be “x” or “y” and a lot of people are searching for “x” and none are searching for “y” it is likely to have been “x.” Taken to the extreme, we might imagine huge stores of information about our activities over long periods of time to produce profiles of our interests. When this is coupled with information about what others are doing, we can begin to produce modern equivalents of 20 year old research efforts at collaborative filtering to produce surrogates and agents who act on our behalf. While we like to think we are unique, the fact of the matter is that when there are millions of us who are similar, we exhibit similar behaviors and these behaviors might be anticipated by agents or surrogates whose goal is to accommodate our needs.

Standardization

While early development of technology is driven by experimentation, there is an ongoing drive toward standardization. This occurs for both technical and economic reasons. Economically, standards help to create larger markets within which the vendors can compete. Technically, particularly in the information technology arena, standards are a requirement for compatibility. So, when technology choice involve standardized and proprietary choices, it is generally the case that the standard choice is the better choice.

Mature standards are so widely accepted that they are almost invisible. The standards for paper and book sizes are good examples. Both come in a variety of sizes but not an endless variety. The effect is that equipment ranging from printing presses to book shelves can be produced efficiently because there is limited variation to accommodate. Information technology is heading toward mature standards. Information technology related efforts are producing a disproportionate percentage -- about 50% of new standards pages – but this is understandable given the state of the technology.

Some recent standards provide a perspective on the importance of standardization. Many will remember the fight between Betamax and VHS. It didn't matter which was better, it just mattered that the public knew there was a standard that warranted their investment. More recently, we witnessed the struggle between Blu-Ray and HD-DVD. The standardization of serial interfaces – USB and wireless Ethernet standards have been even more important. We are still struggling with encryption standards and cell phone standards.

An important area of standardization yet to be resolved relates to documents. We have several standards, both proprietary and public standards, but nowhere near what we need. Chapter 6 will present more about efforts to develop standards for digital documents.

Objectification

In the 1960, object orientation was a paradigm shift in how systems were designed and computers are programmed to manage data and processes. The shift was from thinking about processes to thinking about objects. One story about the beginning of object oriented programming comes from Alan Kay who wrote in a memo to Stefan Ram:[15] "Historically, it's worth looking at the USAF Burroughs 220 file system (that I described in the Smalltalk history), the early work of Doug Ross at MIT … in which he advocated embedding procedure pointers in data structures." Object oriented programming draws from work at MIT on Algol and LISP, in Oslo on Simula and at the University of Utah on Ivan Sutherland's Sketchpad. The story of the US Air Force programming of the Burroughs 220, which Kay worked on, is interesting because it is so simple. The 220 used a set of small tape drives which provided access to data. The Air Force programmer included programs at the beginning of each tape that would extract any arbitrary data that was placed on the tape. This meant that you did not need to know what was on the tape. You simply needed to know that there were a set of programs that could extract or add data to the tape. Consider that in the future, we could trust that any data store would be able to help us store and retrieve the data stored on it without our knowing the details of how it was stored.

---

[15] Kay, Alan, (2003) Meaning of "Object-Oriented Programming. Retrieved March 2011 from http://www.possibility.com/wiki/index.php?title=AlanKayOnDefinitionOfObjectOriented

This shift to object oriented thinking, from procedural orientation, is significant in two ways. First, related to the development of systems, it is a recognition of the need to move to a new more contained modular view of software. Object oriented programming encourages more modularity than traditional approaches. This provides developers with productivity benefits in developing, testing and maintaining the software. As the cost of software development rises dramatically with the increasing complexity of the software, there is increased emphasis on programmer productivity. Object orientation appears to some to be a means by which we can control this process.

The second impact of objectification relates to the user. Traditionally, data (text, graphics, audio, etc.) were stored independently from the programs that processed the data. One consequence of this separation was that the user had to keep track of the programs that processed data files. For example, there are a dozen different formats for storing PC graphics data. The user had to know which program to use. In the object oriented approach, the objects contain both the data and "methods" which are the processes that can act on that data. For example, a graphic object might have a 'print' method that would render the data in a printable form. Another graphic object with a different format would also have a "print" method. Now the user only has to know to invoke the "print" method and no longer cares about the details of the storage format. This provides the user with a simpler model of the system.

Integration

Integration is largely the result of technology maturation. For example, maturing Digital Signal Processing (DSP) technology allows the integration of modem, fax, sound and computer telephony into one device. Integration is manifested through multi-functionality, connectivity, and interoperability. Multi-functionality is being able to use one advanced item instead of a series of less advanced objects. Many office desks once had four terminals on them — telephone, fax, personal computer and workstation or main frame terminal. Now one device can easily handle all these functions – even if most of us keep a separate phone.

In the area of peripherals, we used to have low resolution black and white printers, optical copiers, high resolution color printers, scanners, and fax machines. All of these are now available in a single multifunction device. A scanner plus a printer forms a digital copier. A copier with a modem is a fax machine. There are several commercial products that combine these functions. Performance and quality improvements and price reductions have made a combined personal copier, fax printer, scanner with a price tag of a couple hundred dollars a reality. Similarly, there is a level of integration occurring related to displays. The same screen that displays computer windows is also capable of displaying a television signal. While viewing TV on computer screens has been with us for a couple years, seamless integration of computers with TV's is still just around the corner.

One of the biggest areas of integration occurs related to complex data formats. At the simplest level, it is now possible to include all sorts of objects in a document that used to be limited to text. Not only can I include images, but in digital form, I can include spreadsheets with formula, videos, and animations. A large piece of this integration has been led by Apple and Microsoft and while the formats are not standardized, particularly in the case of Microsoft, they are so widespread as to be widely shareable.

Closely related to integration is interoperability. While there is every reason to believe that the trend of increased interoperability will parallel increased integration, the interoperability trend is not simple or clean. The World Wide Web (WWW) both increased interoperability and served to provide a setback. In terms of document interchange, the WWW provided a significant boost in providing a simple yet rich format that could be produced on one system and displayed on another. At the same time, the semantic of these exchanged documents were not defined. To achieve high productivity, it is important to connect organizations at the enterprise level. Connections between organizations require shared semantics. In 1979, The American National Standards Institute founded a committee – X12 – to develop and maintain standards for Electronic Data Interchange (EDI). EDI was concerned with business transactions such as purchase orders and invoices. It would also be extended to the full range of documents required for the conduct of business across national borders. As should be clear from the date of the beginning of standardization, EDI predates the WWW and relied on a variety of mechanisms for the exchange of EDI data, which was constructed to be efficient for transport over relatively slow network connections. The emergence of the World Wide Web provided a means for exchanging data between organizations that was not standardized. In addition, the emergence of the Extensible Markup Language (XML) provided an alternative and more descriptive means of describing transactions. We will talk more about this ongoing development in a later chapter on business issues.

True interoperability goes beyond connectivity and data interchange and addresses cooperation between programs running on different computers. For example, a user on one computer might invoke a function that causes a program to run on another computer and have the second computer display the results on the first. This aspect of integration has become significant enough for it to be added as a separate trend. Distribution sounds as if it might be the opposite of integration, but in reality it looks to the realization of a vision set forth in 1938 by H.G. Wells in his essays in the book “World Brain.” While Wells anticipated a form of authoritative encyclopedia we now find emerging on the Web, he failed to see the evolution of an immense network of distributed computing.

Distribution

Distributed services, or distributed systems more generally, have been around since networking began. In Chapter 2 (see page 8) on the evolution of the Internet we traced the evolution of key protocols. For me, the protocols that had the greatest impact on the early development of distributed services were the Internet Protocol (IP) and the Transmission Control Protocol (TCP). The Internet (Network of Networks) has been in existence for more than a decade, but there were different schemes for addressing locations on the network. Keep in mind, what existed in the early design was a commitment to exchanging messages as a series of packets that would be routed across the network. The source, destination, and intermediate machines were defined in multiple ways. The Internet Protocol provided what would become a single standard addressing scheme. The Transmission Control Protocol made another equally important contribution. Keep in mind that the base innovation of a packet switched network is the lack of call set up. The model that existed prior to the 1960’s was the telephone model. If I wanted to send a message to a friend, the first thing that I needed to do was to ask the operator to make a connection – to set up a call – between the two parties. This required, in the early days, a physical circuit from party A to party B to be set up before we could talk. In addition, once the circuit was set up, it stayed in place whether we talked or not. Call set up costs and dedicated circuits are expensive, especially for short messages. Packet switched networks simply break a message down into a series of packets and send them into the network to find their way to the destination address. How the routing occurs we will leave as a mystery for now, but it is not much more difficult than having each node know where to send a message next to get to a given destination – basically the nodes don’t know how to get to the destination, only how to get it closer. The problems of packet switched networks relate to lost packets, packets arriving out of sequence, and duplicate packets. These problems are solved by the TCP which provides standard software which interfaces between the programs sending a receiving messages and the network. In the early days we had to write our own code to monitor the process of exchanging messages. Now we can rely on TCP to handle this onerous task.

Many additional problems had to be addressed in the development of distributed systems. Early on, developers were concerned about establishing standards for data that were independent of the many diverse operating systems and hardware architectures. This involved developing both abstract data types and wire formats for the data. Put most simply, communicating systems had to have a common understanding of what an integer was – i.e. 16 bit or 32 bit and of how the integer would be placed in the communication channel – i.e. least significant bit first or most significant bit first. These issues existed not only for primitive data types – integers, floats, Booleans, etc. but for arbitrary data types constructed for particular purposes – i.e. what is a student record or a customer order. There were a variety of different solutions to this problem. The X.400 mail service developed an Abstract Syntax Notation (ASN.1) and Basic Encoding Rules (BER) that served to specify the abstract data types and wire formats to be used by communicating machines. Sun developed its eXternal Data Representation or XDR to serve a similar purpose. With the growing popularity of PC's and the Intel platform, a lot of work could be done assuming common platforms, ignorant of the underlying problem. The emergence of Java and the Java Virtual Machines(JVM) running on multiple platforms provided yet another solution. The Java data types were standard and the JVM provided the interface to the host operating system and hardware formats, not only for communicating processes but for applications in general. Java provided a write once run anywhere solution that included communicating distributed systems. Most recently, XML has emerged as a solution for defining abstract data types using simple character data encapsulated in XML using both the base XML standard and the extensions that allow for strong data typing.

Beyond the issue of protocols and data types, client server systems had to solve other problems. Two difficult problems were the complexity of the code and the problem of rendezvous. Early applications had to know in advance both the location of the server and the port the socket was connected to. The problems of port location and code complexity were both addressed with the development of Remote Procedure Call (RPC) programming. Built on top of XDR, RPC provided an early approach to an Interface Definition Language (IDL). In this case, the interface is the program to program interface. A utility, called rpcgen, translated a definition which included both the remote procedures and the abstract input to and output from those procedures into a series of code fragments. The code fragments included a common set of header files, client side calls, and a skeleton of the server. In essence RPC provides a mechanism for overcoming the data representation problem and provides a structure within which the programmer can think of the distributed system as a set of procedure calls, which happens to be remote. Finally, RPC made use of a standard port to be used in seeking a server side set of procedures allowing the server to be moved from one port to another on a machine without the client having to know about the move. This same functionality was provided in Java with the added advantage that the interface definition language was simply a class defined in java, further reducing the learning curve.

Distributed systems still needed to know what machine the remote process was located on, and the issues of writing the distributed application was still far more complex than normal applications. These problems have solutions although the implementation of the solutions is not yet widespread. There are two parts to the problem again – one is finding a service and the other is programming the distributed application. In the late 1980s, the Object Management Group began work on the Common Object Request Broker Architecture (CORBA). CORBA provides both a way to find a service that is not located on a particular machine and simplifies coding by providing a set of standardized services e.g. transaction management. There are a variety of different efforts that seek to simplify the process of coding and using distributed services on a more ad hoc basis. The OMG and related efforts went a long way to providing both theoretical and proof of concept solutions. In the second section of this chapter, we will talk a little more about how the WWW provided a new form of distributed system.

Over the coming years, distribution will emerge in several forms. They all build on the foundations described above, but they will be introduced to end users using many fancy names. These include GRID computing and Software as a Service (SaaS). While the various forms will develop and emerge over the years, GRID computing generally speaking is an effort to provide supercomputer like services in which a problem is broken down into a series of computational problems which can be distributed across the internet to a large number of machines that can compute results in parallel and combine the results. The vision here is using millions of nodes to do partial calculations that can then be combined. In contrast, SaaS solutions look to use the internet to provide calculations remotely, not necessarily in parallel, that relieve organizations of maintaining select resources as a total system. Rather, an organization can outsource the service and buy it as needed. As a simple example, consider an organization that wants to have a presence on the web, but doesn't wish to buy the hardware, high-speed network connection, and the software for a web presence. A service provider might buy the hardware, network connection, and necessary software, and distribute the cost of the facility over a few dozen or hundreds of customers.

The Technology of the World Wide Web: A Simplified Perspective

The beauty of Tim Berners-Lee's vision of the World Wide Web, which we call the "Web" for brevity in this section, is that it was incredibly simple and extensible by design. Further, it was decentralized and proved fault tolerant. I begin with the most simplistic description of what Berners-Lee wrought and then describe, in a "Mr. Wizard Fractured History" form, how the Web evolved to the form we know today. I think it is useful to begin with the vision of the Web in Berners-Lee's won words:

> *The dream behind the Web is of a common information space in which we communicate by sharing information.*[16]

The Basic Web

The basic model for the web is based on three things: a mechanism for identifying information resources across the internet; a protocol for requesting resources and responding to those requests; a device independent format for documents on the web. Note that I have used the term resource rather than document in the first two points. While documents represent, in my humble opinion, the penultimate container for sharing information, there are many other containers for information. When we discuss documents in the next chapter, we will say more about this, but for now we will stipulate that an image is not a document, but it surely is an information resource. Similarly, I will hold that the performance of a song or a movie is not a document, but it is an information resource. We begin then with an examination of the three basic insights that led to the World Wide Web.

A Unique Name for Each Information Resource

First, and perhaps most important, is the conceptualization of a scheme to reference any piece of information in the shared information space. There was some debate about what this scheme should consist of, and while the resulting standard was far from perfect, it was a huge leap forward. Berners-Lee wanted to develop a scheme that would serve as a Universal Document Identifier. For a variety of reasons, the standardization authority, the IETF, questioned whether or not this was appropriate. There was a debate about names (Uniform Resource Names) versus locators (Uniform Resource Locators) versus Universal Identifiers and other forms. Berners-Lee finally developed a standard (RFC 1738) for a Uniform Resource Locator – "the syntax and semantics of formalized information for location and access of resources via the Internet."[17] It was now possible to build a string that would uniquely identify a resource available on the internet. It should be clear that this is not a universal document identifier –

[16] Berners-Lee, Tim (1998) The World Wide Web: A very short personal history. Retrieved May 2011 from http://www.w3.org/People/Berners-Lee/ShortHistory.

[17] Berner-Lee, Tim et. al. (1994) Uniform Resource Locators (URL) retrieved May 2011 from the IETF website http://tools.ietf.org/html/rfc1738.

capable of uniquely identify every document in existence, but it would serve more than adequately for the vast array of information resources that would soon be placed on the internet. The scheme was incredibly simple, and it would evolve to a more or less final form in RFC 3986. The basic model specifies a scheme specification + a scheme specific specification. Put more simply, you could use an ftp scheme, an http scheme, or one of about eight others. For our purposes, we are going to talk only about the http scheme. The http scheme began with http: and went on as follows:

http://<host>:<port>/<path>?<searchpart>

The host was the internet address – most commonly shown as the human readable name of the address which is provided by a system called the Domain Name Service (DNS). The port was the port on the machine to be contacted for service. By default http services are provided over port 80, so this number appears only if the port is a non standard port. The path is an identifier unique to the host being contacted, but in early use it was a relative directory structure on a Unix host. The search part was provided as a mechanism to locate some particular part of the resource. We will ignore the complexity of RFC 3986, which runs about 60 pages. While the specifics of the standard are required for programming, you have a good sense of both the intent and actual structure of a URI. Put in terms of a simple example, we might find the page which I keep at Pitt using the following URI consisting of the scheme, the host, and a path:

http://www.sis.pitt.edu/~spring/index.html

A Procedure for Accessing Information

With a way of identifying an information resource on the internet, we need a protocol that allows us to access it. RFC 2616, completed in 1999 runs close to 200 pages, and it does not stand alone as a description of the protocol – other RFC's include authentication rules, state management, and other aspects, but for our purposes, the protocol can be accurately described in just a few words. First, http was designed to be an idempotent protocol. That's a fancy word that describes a protocol where each request for service is considered to be complete in and of itself – no state information is maintained on the server. In old fashioned client-server terms, this means that a request is made to the server, and response is formulated and sent, and the connection between the client and the server is closed. The period at the end of the last sentence is meant to be a capital period. When the response is completed it is as if no connection ever existed – at least as far as the early model was concerned. Ok, let's talk about the protocol. Again, I am going to simplify everything while telling the complete truth in what I do say. Under the first version of the protocol, requests could be one of seven types. You could PUT an information resource, GET one, DELETE one, etc. We will use the GET for an example. The first thing we would do is make a connection to port 80 of the server we wished to talk to. We would then request the resource we wanted from that server. In reality, we politely send a fair amount of information in our request, but to get my page above, once connected to port 80 of "www.sis.pitt.edu", we would make a simple request for the resource by sending the following line of text:

GET /~spring/index.html HTTP/1.1

To be correct, I need to indicate that the 1.1 at the end of this line would be followed by 4 characters, a carriage return, followed by a line feed, followed by another carriage return and line feed. I regret the need for that technical aside, but I don't want to hear from the students I tormented with understanding the technical details. When the server knows I am done, it looks for the resource using whatever means are appropriate. (While the sis server is actually using a simple file structure to find the resources, the standard does not say anything about what happens here. This is one of the beauties of the web.) Now it is pretty easy to imagine at least two situations. The server might find the resource, or it might not find it. The reality is that there are many more outcomes, which are represented by status codes. We again turn to a minimal message which would be something like the following:

HTTP/1.1 200 OK
Content-Type: text/html

<html><head><title>MBSpring</title><head>
<body>hello</body>

The message begins with the protocol version followed by that status code where "200"means everything is OK and then a text equivalent that is dependent on the server. In reality, and courteously, the server would have included some more information, followed by a line that says what kind of content is being returned, followed by the carriage return line feed sequence which we show here as a blank line. Then we would send the information resource and close the connection when done – letting the client know they have everything. Actually, as a courtesy we would have sent information about the size of the message and other things. But, most basically, the http protocol is a simple request and response that is most often of the form, "please get this resource for me" followed by "here is the resource you requested."

A Language for Representing Resources

This brings us to the third part of the contribution made by Tim Berners-Lee – the HyperText Markup Language (HTML). Basically, given the fact that the information resources that would be stored and displayed on a variety of different systems, a system independent language was needed for formatting documents. No proprietary format, and there were many in the early 1990's would be sufficient. In the next chapter on documents, we will explore the radical developments that have occurred related to document technology. In that discussion, from a broader document processing perspective, we will be professionally critical of some aspects of HTML. For the time being, we will observe that Berners-Lee correctly observed that there was a way of describing documents, structural markup, that was vastly superior to other existing methods. At CERN, where he developed the Web, the researchers used a way of preparing their manuscripts that was based on the Standard Generalized Markup Language (SGML). SGML provides the syntax for defining classes of document – e.g. manuals, text book, letters, etc. SGML says that the document is constructed as a set of nested elements, that begin with an opening tag (like <author>) and end with a closing tag that matches the opening tag except that it begins with </ (like </author>). Keep in mind that elements consist of opening and closing tags that encompass content which may be other elements of text. Berners-Lee defined a document type that he thought would be good for defining documents and that would include the ability to define a way to reference other documents using the URI we described earlier. For our simple purpose, his definition of HTML allowed us to create a document that included six parts – identify the document as an html document -- <html>, identify two parts, one for information about the document -- <head> and one for the document -- <body> and a title for the resource -- <title> and a link to another resource -- <a> standing for anchor. This given us a simple document:

```
    <html>
  <head>
    <title>
      This is about something simple
    </title>
  </head>
  <body>
    <a href = http://www.sis.pitt.edu/~spring/index.html>
      Here is a link to spring's homepage
    </a>
  </body>
</html>
```

There is obviously more to HTML, such as various levels of headings, lists, paragraph types, etc. They were all simple and intuitive enough to convert millions of users to "structural copymarkers." This was the brilliance of his design of HTML.

The Evolution of the Web

The evolution of the Web can be viewed from two perspectives. The first is technical. It relates to making corrections and amendments to the web to accommodate weaknesses in the original design. The second perspective has to do with the changing nature of people's use of the web. While the vision of Berners-Lee of the web as a common information space in which we communicate by sharing information has proved amazingly accurate, the web as a social network represents a very different information space than the one initially developed. We begin with an examination of the technical changes.

Technical Evolution of the System

The growth in the popularity of the World Wide Web has resulted in many modifications to the overall system. These have been directed at overcoming one or another of the limitations of the system. The Web server and the browser we have described so far might be described graphically as follows:

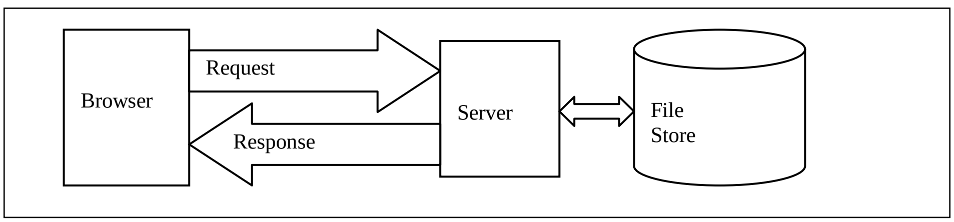


Figure 15: Basic Model of Operation on the Web

The earliest version of HTML that I can find is the June 1993 Internet Draft entitled "Hypertext Markup Language (HTML): A Representation of Textual Information and MetaInformation for Retrieval and Interchange."[18] It is of interest to the current discussion that HTML had all of the elements we know today with the exception of the form element and this is not surprising in that the very first implementations served static documents as we have described above. But very quickly, it became clear that we would want to do searches, as Berners-Lee had envisioned from the outset, and other kinds of server side processing of requests. Beyond searching, we might want to select different documents based on some input or ultimately compose completely dynamic documents. Some of the best work on web servers and browsers was going on at the National Center for Supercomputer Applications (NCSA) at the University

[18] Berners-Lee, Tim & Connolly, Daniel (1993). Hypertext Markup Language (HTML), Retrieved May 2011 from the IETF website http://www.w3.org/MarkUp/draft-ietf-iiir-html-01.txt

of Illinois at Urbana Champagne. They produced the Mosaic browser that later evolved into the Netscape browser. They were also responsible for developing the scheme that became the model for the production of dynamic web pages based on programs. The scheme was called the Common Gateway Interface

The initial HTML draft became obsolete in 1994 and the first formal HTML specification was published as RFC 1866 in November, 1995.[19] By that time, the form element and the various input elements had been added. The earliest reference that I can find to the server side "Common Gateway Interface" (CGI) is a July 1995 document that references several other documents that used to be on the NCSA website, but are no longer there.[20] The formal definition of the Common Gateway Interface was not approved until version 1.1 was approved by the IETF in October of 2004. In reality, dynamic pages had been produced as early as 1995. The basic process was simple. Keep in mind that the actual workings of the protocol are actually somewhat more complex, but conceptually no different than what is described here. The state of the various input elements in the form (the values in text boxes, the radio button that were pushed, the check boxes that were checked, and the elements of lists that were selected) were passed as a part of the request to the server, either as the search component of the URI or as data in the body of a message. The server was responsible for taking this information and placing it in memory – a structure called an environment variable. The server then "forked" a process that was identified in the request. Put most simply in the initial form, the process was an interpreted program, frequently in a language called Perl in early adoptions. "Forking" is a process by which a program makes a copy of itself. Once the process is copied in memory, the operating system continues to run the original process, but also runs the new process which is frequently directed to complete some other action. In the case of CGI, the forked process reads the information placed in the environment variables, follows some logic, and prints some output. That output is captured by the Web server and returned as a response to the original request. The output statement that we programmed made up a complete web page. Conceptually, our original graphic can now be expanded as shown in Figure 16.

---

[19] Berners-Lee, Tim & Connolly, Daniel (1995). Hypertext Markup Language -2.0. Retrieved May 2011 from the IETF website: http://tools.ietf.org/html/rfc1866

[20] World Wide Web Consortium, (1995) W3C httpd CGI/1.1 Script Support. Retrieved May 2011 from the W3C website: http://www.w3.org/Daemon/User/CGI/Overview.html

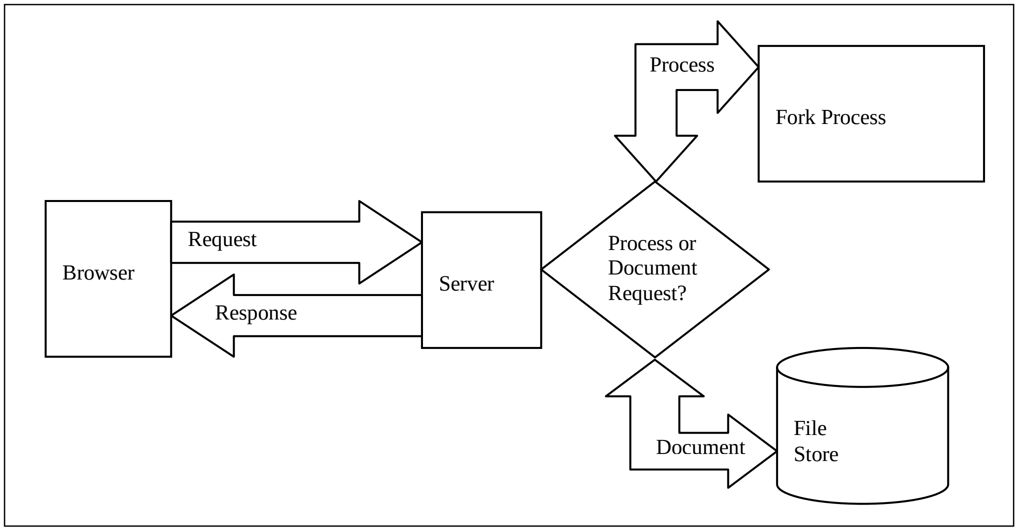


Figure 16: Web Engine with CGI Process

There have been a number of improvements in the basic model as well as the processing of dynamic requests. These include making the process faster by replacing the use of interpreted scripts with compiled program which are faster. In addition, the server side efforts have been made simpler by developing systems where the program that produced a page was replaced by a page that had special scripting elements that allowed it to be dynamic. "Page" had processing instructions embedded. The http protocol was also revised and expanded to add new features. Again the details of these developments are less important than the overall goal of improving the process of access a shared information store. The process of making a request that included multiple objects was formalized and made the standard in http 1.1, with improvements in approaches to compression, caching, and range requests.

One of the shortcomings of the robust protocol developed for the Web is the lack of state information on the server. For a variety of reasons, it would be useful to save information about a visitor to a website, particularly as we move toward commercial uses of the web. The development of a mechanism to manage state information goes all the way back to 1994 when it was first discussed as a solution to a problem faced in developing an application for a client of Netscape. Known as cookies, the idea was that the server would develop information that would be passed back to the client which would store it and return it to the server with subsequent requests. You can see any cookies stored on your computer by asking your browser to show them. You can also instruct the browser not to save cookies if you feel it is a security risk. Today, cookies are generally encrypted as keys to information that is stored securely on the web server.

At the same time, their use remains both controversial and important to the maintenance of information about the relationship between a given browser and a server. The first RFC on cookies was published in RFC 2109 in February, 1997.[21] The most recent version is RFC 2965 published in October 2000.[22]

As forms and dynamic page composition became more common, problems related to incomplete form completion began to surface. Some will recall the early days when a form with an error was returned blank with a request to recomplete the form because some of the data was incorrect. These occurrences were most often due to lazy programmers who found it easier to return a blank form than to return a partially completed form with an indication of where the error or missing data occurred. The solutions to this situation were twofold, first have the programmers take the extra time to provide a reasonable response. A yet better solution relied on this process, plus prevention of bad data and local correction of errors before submission of the request. This was the role of JavaScript, or ECMAScript. It became possible to control what people entered and even to use powerful pattern matching techniques to convert a variety of legitimate inputs to a standard form. The development of local data validation via scripting languages was demonstrated by Sun in late 1995 and was submitted to ECMA in late 1996. The developments were complicated by having different forms of JavaScript for different browsers. None-the-less, the power of the language continued to grow and in conjunction with the development of the Document Object Model ultimately led to the ability to trigger events which contacted a server via a background process and had the ability not just to modify or test user inputs, but to change the basic structure of the page being viewed in the browser. This development was critical to the emergence of Web 2.0 which is described below.

As business applications became more prevalent, security concerns grew. The ability to secure interactions on the Web is both easy and hard. The easy part relates to the interception of messages while in transit on the internet. If we begin with the basic assumption that the endpoints of a communication are safe, which we will show to be not true below, the solution is relatively simple. First, we will develop a method of validating the partners in the interaction. This occurs using certificates that are exchanged between the partners in an interaction. We discussed public-

[21] Kristol, David & Montulli, Lou (1997). HTTP State Management Mechanism. Retrieved May 2011 from the IETF website: http://tools.ietf.org/html/rfc2109
[22] Kristol, David & Montulli, Lou (2000). HTTP State Management Mechanism. Retrieved May 2011 from the IETF website http://tools.ietf.org/html/rfc2965

private key encryption previously (see page 42) and we will return to the topic in Chapter 8. Basically, a server applies to a "trusted third party" who will given them a certificate that verifies they are who they say they are. These trusted third parties register with the various browser vendors and the provision of one of their certificates which can be ascertained using public-private keys allows us to be confident that we are speaking to a given bank. Once we trust the connection, we exchange information that allows both parties to establish a secret key that they can use in communication. Both parties use that key to efficiently communicate across the internet via a protocol called Secure Socket Layer (SSL). Basically, each message passed between the browser and server is encrypted to a sufficient degree to allow us to feel comfortable that the information will not be compromised. Unless we are the Federal Reserve Bank or the CIA, this form of security should offer sufficient protection. Secure Hypertext Transfer was introduced in RFC 2660 in August of 1999.[23]

When we started, we said that we were making some dangerous assumptions. A few of the more obvious clarifications include the following. First, these techniques are fairly robust, but certain kinds of attacks are still possible. A fraudulent "trusted third party" could allow an imposter to circumvent this security. Similarly, there are some other mechanisms that can be used to put a "man-in-the-middle" to intercept communications. A much more significant weakness exists at the two endpoints. The most common is the same one that exists in the non-digital world. A person may observe you typing in your password. It is still secure going over the internet, but it can be captured by someone watching you type. Or you could write it on a post-it note attached to your PC, or you could use one that is easy to guess, etc. Apart from social engineering, you may leave information on your PC and your PC may be compromised using any of a number of mechanisms. Your communications with the bank may be secure, but the password you allowed your browser to store to make your life easier may not be. We have met the enemy and he is us! The second major weakness, as you might have guessed, exists at the other endpoint. While your bank likely does a much better job than the average individual at securing your information, it is still the case that there are humans at the other end that manage that information. So, securing a communication across the internet ends up being pretty easy if we use a little care. It is much more difficult to insure that the people at the end points, particularly the individual user, take all the appropriate steps to insure security.

[23] Rescorla, Eric & Schiffman, Allan (1999) The Secure HyperText Transfer Protocol. Retrieved May 2011 from the IETF website: http://tools.ietf.org/html/rfc2660

There have been many other developments and extensions of the Web that involve making both servers and browsers more powerful. Some of these grew and became widespread standards – such as tabs in browsers. Others withered and died – such as browser side programs known as applets.

The Evolution of HTML

In the next Chapter, we will examine how document technology has changed over the last 25 years. The changes, many of which may be invisible to you, are significant. These developments predate the emergence of the Web, i.e., they are larger than what we see on the web. At the same time, the evolution of the artifacts that constitute the primary mechanism for sharing information on the web are worthy of some consideration here.

The earliest version of HTML was very primitive. The development of HTML1.0, like many things that might be considered incunabula, is lost in the mists of history. Dave Raggett[24] puts the initial prototype in the 1992 period. Dan Connolly's change log[25] indicates that version 1.2 was created on December 3, 1992. The reason for this minute attention to the beginning is not so much to find the actual date, but rather to identify when forms were added. The change log indicates that this occurred on April 7, 1994, about a year and a half later. The period from 1992 through 1995 was a period in which there were many cooks in the kitchen suggesting improvement and additions to HTML based on the various browsers that were being developed. Beyond forms, these included images and other elements found lacking, as well as an increasing number of attributes. (An attribute is considered to be something about an element versus the content of an element. We will say more about this in the next chapter.)

In November, 1995, RFC 1866 was published as a proposed standard for HTML 2.0. It represented an effort by Dan Connolly and Tim Berners-Lee to bring together all of the elements that different browser developers were including and create a single standard. The goal here was to have a standard set of features that would work in any of the many browsers that were emerging. 2.0 became the closest thing to a standard for HTML for a little over a year. While Dan Connolly was working on standardizing a set of components, Dave Raggett was busy working on ideas to extend HTML, like tables, Math equations, more complicated text flows., etc. So at the very time HTML 2.0 was working toward

---

[24] Raggett, Dave et. al. (1998) **Raggett on HTML 4 (2nd Edition)** Addison Wesley Longman, page 17.
[25] World Wide Web Consortium (1995) Markup Change Log. http://www.w3.org/MarkUp/html-spec/ChangeLog. Retrieved May 2011 from the W3C website:

standardization efforts were underway to extend it. Initially, called HTML+ it would emerge in April of 1995, note that this was actually before HTML 2.0. The difference is that 2.0 was submitted as a proposed standard, whereas 3.0 was submitted as a draft for consideration.

In January, 1997 several of the ideas put forward by Dave Raggett appeared as a World Wide Web Consortium (W3C) recommendation. (About this time, responsibility for standards efforts was migrating from the Internet Engineering Task Force (IETF) to the W3C. HTML 3.2 lacked support for equations which would appear as a separate Math Markup Language. It did introduce tables, text flow around images, and other new tags. There were more than 20 new tags introduced. The more significant included:

<APPLET> - To insert an applet
<CAPTION> - To create a caption for a table
<DIV> - To divide a page into logical sections
<MAP> - To create a client-side image map
<SCRIPT> - To create an in-line script
<TABLE> - To create a table
<TD> - To create a regular cell in a table
<TH> - To create a header cell in a table
<TR> - To create a new row in a table

The development of what was to be the final standard for HTML appeared in December 1997 as HTML 4.0, and finally at HTML 4.01 in December of 1999 and with errata in May of 2001. HTML 4 was the first HTML standard to include frames and framesets – allowing more than one document to be shown in a browser window. Frames had been in use for a number of years, but it was not until HTML 4.0 that the standard referenced them. Some elements were “deprecated” – such as “center.” (In the standards world, a deprecated when it is allowed for historical compatibility but its use is discouraged.) New elements, such as “q” for an inline quotation were added. In general, the design goal was to increase structure tags and decrease appearance tags. In regard to this goal, the development was minimal. HTML 4.0 made many significant improvements to the handling of tables and forms, and while tags like “B” for bold were not eliminated, this version of HTML did set the stage for separating structure from appearance with the use of Cascading Style Sheets for appearance.

In January, 2000 the W3C introduced XHTML. XHTML is made possible by work carried forth by the W3C on the eXtensible Markup Language (XML) standard and published as a W3C recommendation in November, 2008. We will have a lot to say about XML and a series of related standards in the next chapter. For now, we simply introduce the failed effort to bring html in line with a more rigorous standard for building structured documents. Where HTML was forgiving of mistakes by users, XHTML had a more strict set of conditions that would ultimately allow for more automated processing of documents. So, while the standards efforts were directed to working out the complex standards required for future development, users were continuing to experiment with extensions to browsers and the HTML language. This has led the W3C to consider a new group – the Web Hypertext Application Technology Working Group. They have decided that the movement to XML is not particularly important and that the future will hold a new HTML specification.

Put in simplest terms, HTML 5 will seek to bring HTML 4 up to date with the changes on the web. There will be new tags for things like audio and video. While XML will be abandoned, there will be more structured elements added to accomplish some new semantic richness in content. In addition, the benefits of an improved Document Object Model mandated by XML will not be abandoned but incorporated. In addition, the HTML 5 effort will seek to standardize the video and audio players such that all browsers use a single standard form.

The Design Evolution of the Web

In addition to the technical evolution based on user needs, both in terms of providers and consumers, the Web is also evolving based on the desire of those responsible for the Web. The vision, begun by Tim Berners-Lee continues with many more contributors. While these individuals have their own personal visions, it would appear to be possible to suggest that in some small way they constitute a shared vision of where the web might go. Indeed, we alluded to one of these design visions in the mention of XML above.

The Web today is not a monolithic entity. It appears in somewhat different forms based on various visions people have espoused for how it might be used. A lot of what we have described above focuses on the Web as a shared information space – the original web was made up of primarily static documents, or resources, which were organized by individuals and organizations to share information. Generally, they were built as a hierarchical fashion and the documents in that hierarchy were linked into a network. This web includes dynamic and generated components but it is characterized by a structure that is focused on a set of "documents" organized by an individual or organization to provide information. It includes individual sites, portals, and marketplaces. This first generation of the web (Web 1.0) might be characterized as the Resource Web:

In 2001, Berners-Lee, Hendler, and Lassila[26] suggested a new vision for the web – the Semantic Web: This is a vision of the web that includes facilities that allow resources, both static and dynamic, to be manipulated algorithmically as well as by humans. It is generally envisioned as an overlay to other aspects of the Web and makes possible the heavy use of agents to conduct work on behalf of humans. This vision is still alive and well, but the intended design has taken interesting turns. From my perspective, the original design of the Web, using a simplified document description language – HTML, was in part responsible for semantics being lacking. As we will discuss in the next chapter, it is possible to design documents such that they can be processed by programs based on semantics. Indeed the development of XML might have solved much of this problem, but as we have already discussed, the XML effort was largely unappreciated. Another effort to bring semantics back to the Web was the development of the Resource Description Framework (RDF). Basically, RDF would provide a mechanism for associating metadata with traditional resources on the Web. In accord with this framework, specialized languages were developed that would enable a program to reason about resources based on the data available in the collected descriptions that would be developed. This effort also seems to have come up short.

Some progress has been made in creating some structure. Researchers are looking at tables that exist in documents to see if they can find structured data in the tables. In some cases, it is not structured data – e.g. some people use tables as a formatting tool. In other cases it is very regular structured data and table structures are correctly used such that the content might be inferred from elements like column heads. Other developments include microformats (such as VCards) and infoboxes used in Wikipedia. These and other approaches to adding semantics continue to develop. An example of a VCard is shown in Figure 18. A small part of an infobox is shown in Figure 17.

[26] Berners-Lee, Tim, Hendler, James & Lassila, Ora (2001). The Semantic Web. Scientific American. 294(5): 28-37.

```
<div>
        <span class="fn n">
            <span class="given-name">Michael</span>
            <abbr class="additional-name" title="Quinlan">B.</abbr>
            <span class="family-name">Spring</span>
        </span><br />
        <span class="adr">
            <span class="street-address">135 North Bellefield</span><br />
            <span class="locality">Pittsburgh</span>,
            <abbr class="region" title="Pennsylvania">PA</abbr>
            <span class="postal-code">15260</span>
        </span>
</div>
```

Figure 17: VCard

```
<table class="infobox biography vcard">
<tr>  <th><span class="fn">Albert Einstein</span></th>  </tr>

<tr class="">
<th scope="row" style="text-align:left;">Born</th>
<td class="" style="">14 March 1879
<span style="display:none">(<span class="bday">1879-03-14</span>)</span><br/>

<a href="/wiki/Ulm">Ulm</a>,
<a href="/wiki/Kingdom_of_W%C3%BCrttemberg">Kingdom of Württemberg</a>,
<a href="/wiki/German_Empire">German Empire</a></td>
</tr>
<tr class="">
<th scope="row" style="text-align:left;">Died</th>
<td class="" style="">18 April 1955
<span style="display:none">(<span class="dday">1955-04-18</span>)</span>
(aged 76)<br />

<a href="/wiki/Princeton,_New_Jersey" title="Princeton,
New Jersey">Princeton</a>,
<a href="/wiki/New_Jersey">New Jersey</a>, United States</td>
</tr>
<th scope="row" style="text-align:left;">Residence</th>
<td class="" style="">Germany, Italy, Switzerland, United States</td>
</tr>
......
</table>
```

Figure 18: Partial Infobox

Other research is working to infer the structure that human have placed in existing HTML documents. Cafarella[27] describes efforts on two projects, one to convert information found in tables on web pages into a more useful form, e.g. column headings may provide schema like information. The other seeks to get at information on the invisible or hidden web – i.e. data that is held in databases or other stores and only accessed as a result of a query processed through a web form.

The most recent development has come in the form of social networking sites. Consider images uploaded to Flickr. Users can add tags to these terms and do so for a wide variety of reasons. These tags may be

[27] Cafarella, Michael, et. al. (2011). Structured data on the web. Communications of the ACM 54(2), 72-79.

descriptive, or they may be silly or idiosyncratic. Theoretically, we can think of the tags as being information or noise. While actual algorithms are more complex, consider a photo that is tagged by 1000 people. Further consider that there are 200 different tags. Tag X was used by 800 people, tag Y by 600, and all of the other tags were used by only one or two people. (It is never this simple.) In this case, we might conclude that Tag X and Y are information about the image and all of the other tags are just noise. The "wisdom of the crowd" might be used to achieve a part of the goal of the semantic web. We will talk more about the vision of a semantic web on page 140 in Chapter 8.

Another vision of the Web is one that sees it as the mechanism that will fulfill the recurring prediction of "thin" clients with netcentric processing and storage. Some readers will remember back to the era of timesharing when users at 100's of "dumb terminals" shared expensive centralized resources. In many ways, this is a much more efficient use of computing resources. At the same time, as computing resources became cheaper, we could afford to let them sit idle 90% of the time. Today, the cost of computing resources is increasingly bound to the software, data, and updates. Keeping computing resources centralized has numerous benefits. O'Reilly[28] suggested in 2005 that this new kind of web experience might be discussed under the rubric of Web 2.0. Web 2.0 is a vision of resources and applications accessed through web servers, but with a high level of local interactivity and personalization. How this is accomplished is somewhat complex, but imagine that we have a web browser on a device that is very sophisticated and that it is connected to the network via a very high speed connection. It would be possible to load something like Microsoft Word into the browser as quickly as Word might be loaded off a local disk drive. Given developments in the Document Object Model and programmatic management of resources on both servers and clients, it is possible to build such programs. It would also be possible to store all our documents and other information artifacts on the Web, much as is already done with email. While there are some efforts to move to thin clients and a netcentric architecture, the more interesting aspect of this vision of computing is the Interactive Web (Web 2.0), which is characterized by heavy client side processing and the provision of an application like, versus page turning, interface for the end user.

It is worth mentioning that many people are interested in what is variously called the invisible, or deep, or hidden web. This part of the web

[28] O'Reilly, Tim (2005) What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Retrieved on August 16, 2010 from http://oreilly.com/web2/archive/what-is-web-20.html

resides in repositories (databases) that are visible only when query information is provided to the repository by either a human or a programmatic agent. It also includes many other resources that require data provided by the request in order to produce a result. It should be noted that there are many restricted sites that are not visible to the public without authentication.

In 2008, Cormode and Krishnamurthy[29] at Bell Labs provided a compelling analysis that distinguished between sites dedicated to information dissemination and sites providing for social networking. Unfortunately, they chose in their insightful analysis to refer to this social web as “Web 2.0”, which leads to unfortunate confusion with O’Reilly’s interactive web. Here, when we use a number, we will refer to The Social Web as “Web 3.0”: This is a vision of the web, well described by Cormode and Krishnamurthy that views people and their contributions as first class objects. Documents are replaced by people. It is a set of web resources that represents individuals and their contributions. Unlike Web 1.0, it tends toward a network structure of people rather than a hierarchical organization of informational objects. Further, its structure and content are defined collaboratively by those who participate rather than by an individual or organization.

All of these forms of the web coexist today. Indeed, any particular site may be a blend that realizes more than one vision at the same time. It is further possible to subclass these various forms. Both Flickr and Wikipedia belong in the Social Web class, but they are very different kinds of social systems. The goal here is simply to lay out the various efforts to make use of this infrastructure. The drive to expand the resource web to a social web is one that makes sense in the context of a digital information revolution moving us to new and better forms of communication.

[29] Cormode, Grahm & Krishnamurthy, Balachander (2008) Key differences between Web 1.0 and Web 2.0. First Monday, 13(6)

Chapter VI: Documents

Introduction

In Chapter 2, we discussed the origin of writing some 7000 years ago. Much of the early writing related to commerce – accounting for transactions. Undoubtedly, there were also some efforts at storytelling and other record keeping. How writing migrated to books and manuscripts is a subject worthy of study in its own. Where and when documents began to be collected is also an interesting question. Some evidence suggests collections as far back as 700 BCE at Nineveh or even earlier at the Sumerian city of Nippur. For a variety of reasons, we choose to view the library of Alexandria as the first famous library – even though there are references to it being modeled after the library of Aristotle in Athens. The library was created about 300 BCE by Demetrius of Phaleron, a student of Aristotle. It was supported by the Ptolemaic Dynasty, initially by King Ptolemy I.

The structure and organization of the library are lost in the mists. Historians believe that it was both a repository and home to a group of scholars from around the civilized world. They were collected through trips to other cities of learning such as Athens as well as through a policy that took the books off ships using the port of Alexandria and making copies of them. The manuscripts that made up the library looked little like the books we know today. There were no title pages, indexes, bibliographies, etc. Estimates suggest that at the at one point there may have been as many as 500,000 – 750,000 scrolls in the library. They were likely organized by various areas of study, e.g. mathematics, astronomy, medicine, geometry. As the number of manuscripts grew, it is likely that the scholars in residence organized and classified the manuscripts to aid in finding relevant information.

I like to describe this period of organizing knowledge as "bucketology." The scholars who managed the collection at the library likely organized the manuscripts such that similar content was located in groups. Other than the suggestion that the organization was similar to that of the library at Athens, I can find no specific information about the level of detail. Over the next twenty centuries libraries continued to be small and generally private affairs with classification schemes particular to the library, be it personal or institutional. There were "public" libraries, but not in the form we think of them today. It would take the invention of moveable-type mass production printing (Gutenberg) to increase the number of volumes and the number of libraries to the point where formal general classification schemes would be needed. Readers of my generation will recall browsing library collections. Looking up one book and finding it in the library would expose you to many other books about the same subject. Of course, given the fact that books could be located at only one place in the stacks, some relevant information might have been placed elsewhere given other content in the document.

We have selected the library at Alexandria as a starting point for the organization of knowledge. (We are suggesting that archives that existed prior to that date were more structures for preserving commercial or government transactions.) In 300 BCE scrolls were the dominant form of documents – a document in one continuous piece. It had little more than a beginning and an end. By the first century CE, documents had taken on a new form, called "codices." A Codex is a manuscript in the form we are most familiar with. The manuscript is broken up into a series of pages that are bound together and placed between more rigid materials – a cover. By the third century, scrolls were about even with codices and by the sixth century codices had completely replaced scrolls. This change created a random access information store that led itself to the organization of information into parts (chapters), indexes of various sorts and attachment of metadata (author, publisher, date of publication, location of publication, etc.)

The Vision of Associative Organization

Vanevar Bush is a huge political figure in American science. As a young engineering PhD, he joined the National Research Council in 1917 to work on war related science research. After the war, he worked to develop what would become Raytheon and taught at MIT where his research involved analog computers and digital circuits. One of his graduate students was Claude Shannon. Through a variety of positions, Bush managed to have the government create, with him as the chair, the National Defense Research Committee whose goal was to coordinate scientific research with military needs. He was responsible for research on radar, sonar, and nuclear weapons among other things. He lobbied for what would ultimately become the National Science Foundation. In July of 1945, he submitted a report to President Truman, in response to a 1944 request from President Roosevelt, called "Science – The Endless Frontier."[30] But we include him here because of an article he wrote for the Atlantic Monthly that was published in July of 1945.

Entitled "As We May Think"[31] Bush summarized a vision of how technology might aid scientific research. The key aspect of the vision was a device, in his case based on microfilm based technology, that could store all of the documents an individual was interested in and store all the links between them. For Bush, the model of human thinking was based on a network structure of links that could be created and traversed by the scientist with ease. He used the term "memex" to describe the device. It is commonly assumed that memex stood for memory extension. In reading the article, one gets a sense of Bush's commitment to science enhancing the human experience, his lifelong commitment. He imagines the power of science enhanced by a personalized indexing system that vastly increased the access to information he found lacking in traditional classification systems. He also imagined automatic translation of speech and the capture of experiences via cameras where observations could be immediately captured and incorporated.

Bush's view of the organization of knowledge influenced many people. Notable among these individuals were a young computer engineer by the name of Douglas Engelbart and a computer scientist by the name of Tim Berners-Lee. (Apologies to hypertext scholars who would like to see some discussion of the import of Ted Nelson.) Engelbart had read Bush's article in 1945. After serving in World War II he returned to complete his bachelor's degree at Oregon State and then went on to earn his PhD in Electrical Engineering and Computer Science at UC Berkeley in 1955. Engelbart is the inventor of the mouse, and a device called a "chord keyset" which was capable of indicating commands the user wished to exercise on the object pointed to by the mouse. The mouse was used to select an object, and the chord keyset could be used to execute one of thirty one commands.[32] The chord device, became the buttons on mice. As best I can

[30] Bush, Vannevar (1945) Science: The Endless Frontier: A Report to the President. Washington: United States Government Printing Office Retrieved May 2011 from the NSF website: http://www.nsf.gov/about/history/vbush1945.htm

[31] Bush, Vannevar (1945) "As We May Think", *Atlantic Monthly*, July, 1945, pp 101-108.

[32] This is another place where powers of two come into play. With one button, two states are possible – 0 or 1, up or down. With two buttons, four states are possible 00, 01, 10, 11. With five buttons, thirty two states are possible. If we exclude the state where no buttons are pressed, thirty one commands are possible.

tell, Engelbart produced the first digital hypertext system, called NLS (for oN Line System) which was publicly demonstrated in December 1968 at the Fall Joint Computer Conference. Despite these and other equally impressive technical developments, Engelbart's most important contribution may be a paper he wrote on "Augmenting Human Intellect."[33] He proposed a research framework that might be used to better understand the process. He suggested that "the system we want to improve can thus be visualized as a trained human being together with his artifacts, language, and methodology." Aspects of this paper became very controversial in the short run. In the long run, the notion of augmenting human intellect and the process and view adopted by Engelbart were recognized as being of great import.

Engelbart's ground breaking research and his inventions had a great influence on many researchers and research programs. Several members of his research team, frustrated with some of his views left his "Augmentation Research Center" and joined the researchers at Xerox PARC who developed the first graphical personal computers, networked via LAN and connected to laser printers. Others interested in hypertext were inspired by his accomplishments. These included Andries van Dam at Brown University who developed a Hypertext Editing System in 1967, which was used by NASA. Also influenced was a young British computer scientist by the name of Tim Berners-Lee who developed a system called ENQUIRE in 1980. These and many of the early hypertext systems, had typed links and other semantics that made them much more useful than the hypertext we know of today on the Web.

A Slow Return to Structure

The structure of documents in the paper world had emerged over a 2000 year period. Scrolls gave way to codices which provided the infrastructure for random access text, sectioning, indexing, and cross referencing. Citations eventually developed that allowed document to document references. Classification systems emerged that allowed collections to be organized. In some sense all of this was lost with the transition to digital documents.

Computers were originally dedicated to performing calculations. Text processing was not a primary focus, perhaps with the exception of the fact that high level programming languages consisted of lines of code which

[33] Engelbart, Douglas (1962) Augmenting Human Intellect: A Conceptual Framework. Retrieved May 2011 from the Engelbart website: http://www.dougengelbart.org/pubs/augment-3906.html

were compiled into machine code. Most histories of text processing begin with the Unix operating system developed at Bell Labs which had an editors called "ed." While it is likely that editors of some sort might be found as early as the 1950's I would suggest that two can be found in the 1960's. TECO, an acronym for **T**ext(originally Tape) **E**ditor and **CO**rrector) was developed circa 1963 by Daniel L. Murphy at (MIT) around 1962. It was actually both a text editor and a program processor and eventually grew into emacs. QED, an acronym for Quick Editor was a line-oriented computer text editor developed by Butler Lampson and L. Peter Deutsch at Berkley and would migrate to "ed" and then "ex" and then vi at Bell labs in the 1970's. For most computer users, the use of expensive computers to do simple text processing was out of the question. While the descendents of these editors would be used by secretaries during the early years, the editors had been designed as "line" editors that were optimized for programmers who used them to write programs in ways that were exceptionally more convenient than writing the programs on punched cards. These systems had no sense of sentences, paragraphs, chapters or references.

The Alto and the Star

One of the greatest, and in some ways saddest, stories in the history of digital documents emerges from the halls of Xerox's Palo Alto Research Center (PARC). This story has been well documented in books and articles.[34] The story begins with Peter McColough who took over the presidency of Xerox in 1966. Working with the Xerox Chief Scientist, Jack Goldman, McColough believed that Xerox needed to engage in research that would allow Xerox to move beyond copier technologies. Goldman, who had been a director of Scientific Research at Ford knew what he wanted in a research director and suggested a physicist by the name of George Pake who was the Provost of Washington University. Goldman and McColough assured Pake that he would have the time and money to develop a basic research agenda. They charged Pake with inventing the office information architecture of the future. The story of the lab's location and its personnel read like a spy novel.

Basically, they succeeded in inventing the office information architecture, but Xerox, for a variety of reasons failed to capitalize. Some of the accomplishments include:

[34] The business story is perhaps best told in Douglas K. Smith and Robert C. Alexander. **Fumbling the Future**. HarperCollins Publishers, 1989. The technology story is well documented in Michael A. Hiltzik. **Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age**. Harper Business, 2000.

- The development of the WIMP interface – windows, icons, mouse, pointer (1975)
- The development of WYSIWYG[35] text editors (1974)
- Object Oriented Development Environments (Smalltalk) (1972)
- The realization of the Ethernet local area network (1973)
- Documents with integrated text and graphics in multiple languages (1974)
- Laser Printers (1971)

These accomplishments may seem less important unless you think carefully about the dates. The first IBM PC was introduced in August of 1981. These developments took place more than 5 years earlier, and in the case of the laser printer, a decade earlier. Five months before the primitive IBM PC was introduced Xerox brought the Star to market! Obviously it was much more expensive – the one that I was given for research, along with the communications and file servers and the laser printer would have cost around $60,000 at a time in 1981 when the very expensive IBM PC cost about $2000.

But for the majority of users, multi language, WYSIWYG text editing output to a laser printer over a high speed local area network was not in the cards. While the Mac, which was copied from the Star was introduced in 1984, it was a mere shadow of the Star. The Microsoft world did not embrace the WIMP interface until Windows 95 was introduced. (In reality, Windows 3.0 was introduced in 1990 and provided a windows like interface for DOS machines.) Having used all three interfaces, it is clear what Xerox produced in 1980 was superior to any other system well into the 1990's, maybe even through the year 2000. Most of us had to be satisfied with separate software systems for word processing and document formatting.

[35] What You See Is What You Get,

Word Processing and Desktop Publishing

The history of word processing on personal machines is rich and dates back to the mid 1960's with the IBM Selectric Typewriter combined with a magnetic tape cartridge that could store 28,000 characters. In 1969, a magnetic card version of the Selectric was added. Through the 1970's dedicated word processing systems were introduced to the market. Wang Labs and Lanier were notable contenders in this arena. A rich history of this period has been provided by Thomas Haigh.[36]

Most of the early wordprocessing systems actually predate the IBM PC. (WordStar was originally written for CP/M operating system that ran on a variety of microcomputers. WordPerfect was originally written for Data General minicomputers and ported to the IBM PC in 1982.) It is almost painful to go back and look at some of commands and machinations one had to go through to use these systems – control keys, modal editing, etc. With the exception of the Xerox Star, most PC based system output to dot-matrix printers and robot typewriters. For the most part, there was very limited proportional spacing and documents were printed with a single font and those fonts were of limited sizes. Adequate for letters, they were generally unacceptable for publishing. Today's MicroSoft Word comes very close to providing both personal and professional publishing capabilities even though people working on very large documents with multiple indices and cross references will find that some tasks are more difficult to manage than others.

From the mid 1980's through 2000, systems generally fell into two categories – PC-based wordprocessing systems and desktop publishing systems. We take a look at each of these in turn.

Word Processing and Procedural Copymarking

As indicated, the earliest word processors emerged from line editors written by programmers for programming. As office workers began to use these editors for text processing, it was necessary to adjust the model and the mode of operation. Looking back to the command set for WordStar, the following cursor movement commands can be found: "Right character, Left character, Up line, Down line, Right word, Left word, Top of screen, Bottom of screen, Beginning of file, End of file." Nowhere near the ease of pointing somewhere with a mouse, but even short of that, note that the only "text" structure movement is "word." It was not possible to move by

[36] Haigh, Thomas (2006)"Remembering the Office of the Future: The Origins of Word Processing and Office Automation," IEEE Annals of the History of Computing, October-December, 2006 pp. 6-31. Also: http://www.tomandmaria.com/tom/Writing/Annals2006WP.pdf

sentence, paragraph, or page. It may be of interest to some to note that in 1980, the Xerox Star had a convention that a single mouse click selected a point, a second a word, a third a sentence, and a fourth a paragraph. If you asked for "properties" you got the property sheet appropriate to what was selected. If you are using Word, try clicking the mouse multiple times without moving it. I believe you will find word now selects points, words, and paragraphs. All of the command structures had to be migrated from lines and spaces to more text oriented metrics.

If you remember the time when we used typewriters to compose, you probably had a tendency to hit the enter key at the end of each line when you began to use a word processor. Similarly, rather than adjusting tab settings, you probably inserted multiple tabs between words in a "table" to get things to line up. Finally, it is likely that you did, and may still, insert a tab before typing the first line of a paragraph rather than adjusting paragraph settings. Similarly, to adjust the space between paragraphs, you likely added a blank line rather than adjust the inter-paragraph spacing.

The first solution to text processing was the development of procedural markup codes. They were initially very similar to the kind of markup done on traditional documents to prepare them for typesetting. For example, Figure 19shows the "markup" that might be provided by a professional editor. The instructions to the typesetter indicate that this text is to begin on a new right handed page and the title is to be set centered in 40 pt Bold Arial type. The author and organization is to be set centered in 24 pt Italic Arial. The paragraph text is to be set in 10 pt Times Roman with two points of leading between lines.

This approach to copymarking was mimicked in word processing systems. Whether the commands were visible in the editor or hidden by a WYSIWYG interface was pretty much a matter of the state of the development of the editor. Almost all editors eventually became WYSIWYG, but underneath the interface was the same kind of copymarking visible in early editor. The example below is from Peachtext – one of my favorites for pure text editing. It was used to preparing text files for Ventura publisher – described in the next section. The commands come in many different forms.

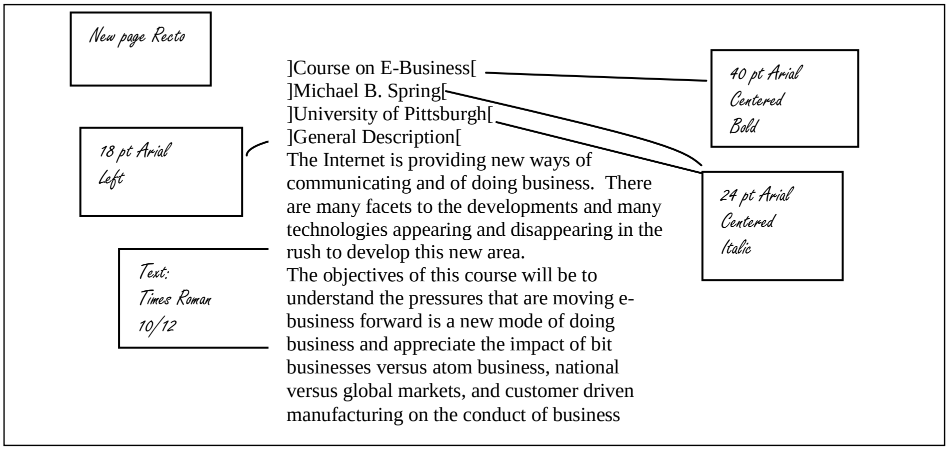


Figure 19: Layout Editing

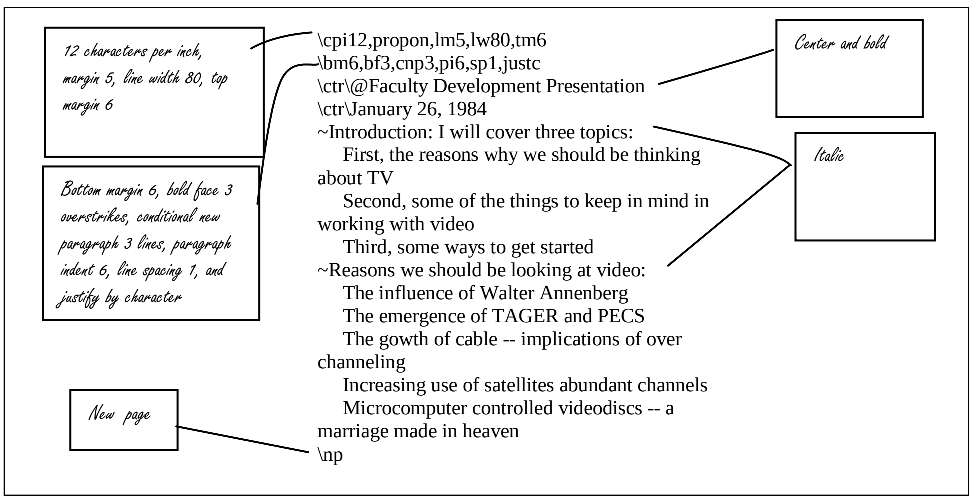


Figure 20: Procedural Copymarking

Some are reserved characters – such as '@' and '~' to provide access top commonly used commands such as bold and italic. Others are global and set at the beginning of the document such as line spacing and the number of lines needed at the bottom of a page before a new paragraph is started – cnp 3. Other commands provide overrides to the global formatting e.g. /ctr/ or a special command such as the command for a new page.

The next, revolutionary step, in the markup of documents occurred in the mini and mainframe computer environment discussed on page 95.

Desktop Publishing

In 1984, following the Apple Mac, Apple introduced the Apple LaserWriter which as capable of 300 dot per inch (DPI) printing. At the same time, based on the same print engine, Hewlett Packard introduced the

HP LaserJet. These printers required a control language and for reasons of compatibility the Printer Control Language (PCL) was chosen. While PCL was a capable language, a much better printer control language called Interpress had been developed at Xerox PARC. Xerox's failure to capitalize on the potential of such a language led more PARC employees to leave and form their own company—in this case Adobe. Adobe released Postscript level 1 in 1984. Postscript was a very special language in the mathematical precision with which it allowed a page to be described. The Mac, postscript, and the LaserWriter made it possible to develop desktop publishing software of some significant power. Aldus, named after Aldus Manutius of italic type fame, released Pagemaker for the Mac in 1985. From a graphic design point of view, this was a killer app and solidified the reputation of the Mac in the graphics community. On the PC side, another frustrated cadre of Xerox PARC employees left and formed Ventura Publisher in 1986. These systems had a significant learning curve, but offered significant control of output.

Pagemaker was focused on intimate control of text and graphics for publications like brochures. Ventura was optimized for long standardized publications. Some of the more interesting capabilities that these systems made available are described below. While all of these things can be accomplished today, sometimes with great difficulty, in the late 1980's, these were beyond the capability of most word processors.[37]

Page controls. It is easy to imagine that we might want to insert a page break, but we might also want that page break to be a new right hand or rector page in a printed publication. Further, we might want to insure that one line of a paragraph was not left at the bottom or top of a page – widow and orphan controls – and we want these controls to be adjustable. Further, we want to be able to say that a section heading should be kept with the text it was a heading to and not left at the bottom of a page. Add to all of these considerations a complex document that has footnotes. It may be that footnotes need to be broken over multiple pages so as to not disrupt the reading page flow – how big can a footnote be and what kinds of page control can be set.

Line Controls. At a crude level, a good hyphenation algorithm was needed that could be turned on and off. Beyond the algorithm, hyphenation dictionaries should be allowed that identified terms to be excluded from hyphenation or that indicated preferred hyphenation points that were different from those that would be suggested by the algorithm. Beyond the

[37] Note that these controls are only a partial list and meant to illustrate some of the more arcane kinds of controls that needed to be programmed.

individual line controls, there are a variety of line spacing controls. Not only interline spacing, which needed to be very precise, but spacing before and after entities and between different kind of entities. For example, a list might be single spaced, but some additional space should be allocated before the first item of after the last. Line spacing can also be taken to the level of wanting to adjust line spacing automatically based on page controls. After all widow and orphan controls have been executed, adjust line spacing within specified limits such that the lines of text are justified vertically.

Spacing controls. There are aesthetic concerns with fonts at different sizes. At the danger of oversimplification, when a word is printed in a larger font, it does not look correct unless some of the space between the letters is removed. Similarly, and again related to graphic design aesthetics, we want to "expand" rather than "condense" inter-character spacing. A specialized consideration in spacing is that particular character pairs do not look right if normal spacing is used.[38] Thus, if we look at the following words – WATER and WATER, we view the second as looking normal. That is because kerning has been done. From a typesetting point of view we have overlapped the type blocks of the second water. Of interest, kerning from the point of view of a designer is something that must be adjusted across font sizes. There are hundreds of such letter combinations.

Indices: There were any number of indexing controls that were required. These systems made it possible to generate the indices based on inline information. There are at least three distinctly different indices that were part of the systems. First, the traditional back of the book index needed to be generated. By placement of a command in the text, it was possible to place a word in the index different from a word on the page referred to. In addition, it was desirable to have multiple levels of indices. For example, some items should be indexed under a main heading. A second form of index includes the various tables of contents – not just a listing of the major textual blocks, but lists of figures and tables and examples. Finally, there are page indices that are placed on left and right handed pages. These might easily identify the page number, but could also include the chapter and section. At the limit, these might include the first and last element on a page – much as a telephone book identifies the address entries on a page.

These are only a few of the functional requirements of desktop publishing systems, but they provide a flavor of the kinds of functionality

[38] The origin of kerning arises with Gutenberg and typesetting. In hot metal typesetting, each character is placed on a block which fixes how close it can be to the other letters, in the case of known kern pairs, special slugs had to be created to allow the letters to look like they were correctly spaced.

that were built into these desktop systems based on what was available in more powerful, and much more expensive professional systems.

Markup, Macros and the Emergence of Structural Copymarking

Where the vast majority of the efforts on personal computers were directed toward the goal of integrated editing and formatting with a WYSIWYG interface, the mini- and main-frame systems tended to completely separate editing and formatting programs. Like the editors, the history of formatting programs is long and rich. There is reasonable support for the RUNOFF program written by Jerome Saltzer for the CTSS operating system. It was ported to Multics by Bob Morris and Doug McIlroy as runoff which was then ported to Unix as a whole series of formatting programs – nroff, troff, gruff, etc. The Unix man pages identify Joe Osanna as the "father" of the series, but the various versions were the product of a number of programmers including Ken Thompson, Brian Kernighan and Dennis Ritchie and Douglas McIlroy, all of whom were involved in the development of the C programming Language and the Unix operating system at Bell labs in the late 1960s. We begin with the most basic of the roff series, nroff, which was written by Osanna.

Nroff commands are of two types. Dot commands are on a line by themselves and the line begins with a period'.'. Slash commands are embedded in text and begin with a '\'. Many of the dot commands take arguments. A few of the many dot commands are shown below.

.ll 7i	Set the line length to 7 inches	.bp	Break page (start a new page)
.ce 2	Center next two lines of text	.br	Line break (start a new line)
.sp 3	Skip three lines (add blank lines)	.sp 1i	Skip down 1 inch
.ps 12	Use 12-point type	.vs 14	Use 14-point vertical spacing
.nf	No fill (stop filling lines to the line length)	.fi	Fill

A couple examples of the slash commands include:

\fB	Start printing in boldface	\fI	Start printing in italics
\fP	Resume previous font style	\s12	Print in 12-point type
\s+2	Increase point size by two	\s0	Resume previous point size

With time, additional commands were added, but the big evolution in formatting occurred first through "packages" of macros and then by the addition of preprocessors. Preprocessors handled higher level concepts, translated into basic commands for tables, equations, and bibliographical references and other objects. Macro packages allowed the user to use a set of commands by a more friendly name. That is, rather than issue the same five commands for every header, a macro was defined that represented the header. This began the movement away from procedural copymarking. One of the earliest of the many macro packages was –ms. A few of the hundreds of –ms macro commands it made possible are shown below.

.NH n	produces a numbered heading where n is the heading level
.SH	produces a section heading
.PP	starts a normal paragraph with first line indented
.LP	starts a normal paragraph with all lines flush left
.IP *label*	starts an indented paragraph with the optional *label* is printed to the left of the first line.
.KS *text*	the .KS command is matched by a .KE command after the text. The text will be moved to the following page if it doesn't fit. A blank space may be produced at the bottom of the page.
.KF *text*	works like .KS, but float *text* to the following page if it doesn't fit on the current page. Following text may be moved forward to fill the bottom of the page.

This review would be remiss without the mention of at least two additional systems. Donald Knuth is a computer science legend. He is most well known for his multi-volume publication "The Art of Computer Programming." What eventually became a quest to complete a seven volume set, of which four have now been completed began with the first volume on *Fundamental Algorithms* published in 1968. The story of the project is worthy of a read in and of itself, but our part of the story continues with the publication of Volume 2 in 1976. Between Volume 1 and Volume 2 the publisher had moved from mechanical typesetting to computer based phototypesetting. He was unhappy with the quality of the typesetting program and decided to take a few months to write a better program. Eight years later he had completed TeX which was to be used to set the type for all his books. Like the other systems we have discussed, TeX was a markup language. It was different in that it focused heavily on mathematical equations and their representation for optical phototypesetters. It also had exceptional sophisticated hyphenation algorithms. While TeX had commands that were more mnemonic than some of the very early formatters, it was still more complex than most users wanted. TeX's macro capabilities were used to build more user friendly versions that were widely used in technical publishing. The original version of LaTeX was written by Leslie Lamport. It is not unusual to find a document produced using LaTeX to this day.

The Xerox Integrated Composition System (XICS) also merits a brief mention. While it has passed into history, in the mid 1980's Xerox was still a power in electronic publishing, notably because they were producing the fastest laser printers. Still among the fastest ever produced, the Xerox 9700 could print 120 pages per minute. Before a world of distributed processing, this was an important component of the system. Like the other systems, XICS had a macro language that allowed generalized copymarks to be designed, and the software was capable of producing optimized output for the high speed laser printers. In 1986, we worked on a research project that produced custom textbooks from a collection of fragments constructed on demand. We worked through the latter part of the decade with Xerox to develop a software system that took documents produced for wordprocessors and using a virtual machine convert them into a document marked up with dynamic XICS macros.[39] The software endeavored to bridge the gap between procedural copymarking and structural copymarking, which will be addressed next – on page 98.

The final system we look at is SCRIPT, the IBM scripting language for text formatting. It looked very similar to the roff family and used dot commands to direct formatting. Like the roff family, SCRIPT achieved real power when it evolved to include macros. In this form, it was known as the Generalized Markup Language or GML. A sense of where macros were evolving can be gleaned from the user guide for GML published in 1980.

Chapter 3. Paragraphs and Headings
Chapter 4. Lists
Chapter 5. Examples and Figures
Chapter 6. Highlighting, Citing, Noting, and Quoting
Chapter 7. Tables
Chapter 8. Cross-References and Footnotes
Chapter 9. Overall Document Structure
Chapter 10. Indexing
Chapter 11. Process-Specific Controls[40]

The nroff macros continue to become more user friendly in how they are described. The example below shows a paragraph with an example embedded in it.

:p.Instructs the system to begin a paragraph. Then in the middle of the paragraph we want to put an example which would naturally be indented.
:xmp.
We might have some code here
:exmp.
:pc.This would then be a continuation of the paragraph above, likely
meaning that it would not have its first line indented

[39] Spring, M.B., (1989) The Origin and Use of Copymarks in Electronic Publishing. Journal of Documentation, 45(2), pp 110-123

[40] IBM Corporation (1991) GML Starter Set User's Guide. Retrieved May 2011 from the IBM website: http://publibfp.boulder.ibm.com/cgi-bin/bookmgr/FRAMESET/dsm04m00/2.0?DT=19910507103243

The macro commands are generally more meaningful to a non technical user. You will also note that for the first time in the examples we have shown that the macros include an indication of where to begin and where to end. This feature will become increasingly important and allows for nested structures such as the lists in this last example.

:ol.
:li.*First item in a first-level list.*
:li.*Second item in a first-level list.*
It has a nested list within it.
:ol.
:li.*First item in a second-level list*
:li.*Second item in a second-level list*
:eol.
:li.*Third item in a first-level list.*
:eol.

Structure Emerges

There are ongoing discussions about exactly when the shift was made from procedural copymarking to descriptive and then structural copymarking. The examples above show the beginning of descriptive copymarking – the shift from point sizes and layout to codes like "ol" for "ordered list" and "li" for "list item" are surely more descriptive. The label "structural" will be reserved for copymarking that is not only descriptive but enforces structural rules – i.e. the document is not only marked up with terms that are more semantically meaningful, but that markup is such that the document may be processed with confidence algorithmically. As with many fundamental changes, there is no absolute and clear cut event that can be pointed to. The basic credit is given to Brian Reid. At the same time, we note that what happened in 1980 is, in some ways, still incomplete as a full model in 2011.

Scribe

Brian Reed was studying for his PhD in Computer Science at Carnegie Mellon University.[41] His dissertation demonstrated the design of a language for describing documents and a compiler for processing those documents. The markup language was written in terms that made sense for an author and the markup was definitively separated from layout. In his

[41] Reid, Brian, (1981) Scribe: A Document Specification Language and its Compiler. Ph.D. Dissertation. Carnegie Mellon University, Pittsburgh, PA, USA. AAI8114634.

system, he made a couple simplifications to the process of marking up text that made great good sense at a time when terminals and network

```
@make(report)
@begin(titlepage)
@title[COMPUTER CENTER REPORT]
@date[January 12, 1984]
@end(titlepage)
@chapter(DEPARTMENTAL LIBRARIES)
The library for Computer Science, CSL:, has been created, with a quota of 10,000 blocks.

Free space on SPL: was critical during the Fall term. It is currently at 106,000 for System
A and 122,000 for System B, and will decrease rapidly as the Winter term progresses.
@section(INFORMAL COURSES)
The schedule of informal courses for the Winter term has been announced.The courses being offered are
@begin(list)
Computing for the New User

Introduction to Graphics at Pitt

Interactive System 1022
@end(list)
Please see SYS:NEWS for details.
```

Figure 21: Document Markup with Scribe

communications were slow. For example, rather than require that the user identify every paragraph, which ultimately we will require for structured copymarking, this most common tag was not required. Text that was not otherwise marked was assumed to represent the most natural child entity, and blank lines were an indication of separation of elements.

The first line specifies that this document will be a report, not a brochure, or a letter, or a thesis. Scribe provided for about a dozen simple and complex document types. It was also possible to add additional document types and to define the components of those documents. Some elements of Scribe documents have start and end markup or tags – e.g. @begin(titlepage) and @end(titlepage) while others place the text in the parenthesis or other delimiters that follow the markup – e.g. @section (Informal Courses). Things like the elements of a list or paragraphs were assumed based on blank lines separating text lines.

Scribe provided a plethora of automation. Tables of Contents and References were handled automatically. Indexing, footnotes, and citations were extremely easy to use and provided high quality results. All of the kinds of advanced features discussed under desktop publishing had been thought of. Hyphenation algorithms and exceptions were well implemented. Scribe was used to format my first two books. The source files are amazingly clean – i.e., there is very little specialized coding. As a matter of fact there is none. The postscript output files produced were simply loaded into the publisher's typesetters. More work for me, but this capability was especially appreciated for my book on Postscript[42] where the system allowed me to generate the postscript images from the postscript code that was shown in the book eliminating the possibility of errors in the code.

The Standards Battle

While Brian Reid and Scribe get credit for establishing the base theory for descriptive and structural markup, the focus would shift away from Scribe to two similar approaches to markup. IBM would move it's Generalized Markup Language to the Standard Generalized Markup language (SGML) which would become a standard in 1986. It would compete with the Office (later changed to Open) Document Architecture (ODA), a standard developed under the auspices of the European Computer Equipment Manufacturers Association. Like many of the ISO and CCITT standards developed with European leads in the 1980, ODA was a powerful and complete standard for a standard document architecture and interchange format. Consisting of more than twelve parts, there was little ambiguity in the standard. SGML on the other hand was more focused on the basics of markup and less focused on the specification of the processing details. Indeed, the various components of the SGML standard needed to complete the standard were never finished. (They were, along with others, completed as a part of the XML family of standards which superseded SGML in 1998.)

What the standards did was to codify several important points about the design of documents optimized for the digital environment. First, and most important, the standards separated the logical structure of a document from its appearance. The logical structure of a document defines the hierarchical organization of the information in a document. It says nothing about the layout or appearance of the information. The layout of a document was to be separately defined and indeed was specified as a part of ODA. So where the logical structure included chapters, sections, subsections, paragraphs, footnotes, etc, the layout structure included pagesets, pages, blocks, etc. Both of these structures were defined as trees with content at the leaf nodes of the trees. In addition to the separation of logical and layout structures, both standards introduced the notion of document types. While ODA introduced a basic initial set of document types, SGML, with the exception of examples, did not define basic document types but simply the rules for defining them. Both standards also

[42] Spring, Michael & Dubin. David (1992) Hands on Postscript. Carmel, IN: Hayden Books.

defined basic content types. Again, SGML indicated that they needed to be defined while ODA defined some of the basic types.

While many observers viewed ODA as the superior standard, most organizations moved to adopt and develop software for SGML. The adoption of either standard was sparse. Organizations that needed to produce large documents of archival value understood the importance of standardization and structure. Thus, it is not surprising that Boeing mandated the use of SGML for its documentation. It is rumored that the documentation for a Boeing 747 equals the weight of the 747! Similarly, the Department of Defense, where the documentation for weapons systems is complex, constantly changing, and active over the multi decade lifespan of the weapons system, adopted the standard. The standard had also been adopted the European Organization for Nuclear Research (CERN) for use in publishing the papers written by the scientific staff. A computer scientist by the name of Tim Berners-Lee, who was working at CERN, was aware of the standard and decided to write a document type for use in his hypertext system.

HTML Revisited

The Web and HTML provided a tremendous boost to the use of descriptive markup and structured documents. At the same time, HTML became an obstacle to good document structuring. In order to get people to use HTML, it was made as simple as possible, and browsers were designed to be very tolerant of sloppy coding. For example, in strict SGML and XML markup, every element must have a begin tag and an end tag. Most browsers were tolerant of missing tags and provided implied end tags. This makes it much more difficult to algorithmically process these documents.

Despite the intent to make markup descriptive of structure, many HTML tags either implied or explicitly defined appearance. In early browsers the appearance of H1 was set. Similarly, tags like “b” for bold, “u” for underline and “center” explicitly specified appearance. Many users, especially early on, used tags indiscriminately to make the text appear as they desired – i.e. they knew that the font for h1 would be bigger than the font for h3.

But the biggest shortcoming of HTML was the lack of imposed structure. For example, most would agree that a book consists of a series of chapters and chapters consist of a series of sections which in turn contain subsections which contain paragraphs and examples and figures. If a section contained a chapter, we would sense some dissonance. If the book did not have a title page, or if it were not at the front, we would sense that something was wrong. HTML was virtually devoid of structure, and where the standard suggested structure, browsers generally were tolerant of the misuse. Thus, the structure problem had three components. First, the document definition had minimal imposed structure. Second, the elements lacked tags like “author”, “publisher”, etc. and the elements that were provided were more about layout than logic. Third, every document put on the web used the same document definition. A thesis, an article, and a poem all used the same set of tags. The shortcomings of HTML were quickly understood but the cost of modifying browsers to handle multiple different document definitions, especially if they were of the form defined for SGML was prohibitive. Therefore, the World Wide Web Consortium gathered talent to rewrite the SGML standard and make it “simpler.” The task began almost immediately and while it is largely complete, it took a long time to produce all the standards needed to define the successor to SGML – the eXtensible Markup Language (XML).

The eXtensible Markup Language (XML)

Providing an overview of XML in a few pages is a daunting task. I have devoted an entire graduate course to introducing the basic concepts. Thus the reader is advised that what follows is an effort to provide a simplified but I believe accurate sense of what it is all about. To get started, we introduce a few pieces of terminology. A document will consist of a set of nodes arranged in a tree structure. These nodes will be of many different types – nine to be exact, but for our purposes we will be concerned primarily with those that define the “elements” and the content of the document. An element will appear in text as follows:

<author>Michael B. Spring</author>

In reality, this defines two nodes, an element node called “author”, and a text node containing the text “Michael B. Spring. For our purposes we will simply refer to the whole string as the element. The element then is made up of a “start tag”, the content, and an “end tag.” If we wanted to provide information about the element that was not intended to be displayed as content, but was important to the element, we would define it as an attribute. An attribute of the author element might be something like whether the author is a corresponding author – the author from a group who serves as the primary contact. Such information could be provided in the markup and would look as follows:

<author corresponding = “true”>Michael B. Spring</author>

Defining a Document Type

The first thing we need is a definition of a class of documents. This was accomplished in SGML through a Document Type Definition (DTD) that included three kinds of definitions, elements, entities, and attributes. In the SGML standard, written during a period or dumb terminals, entities served several purposes such as reducing the amount of typing by defining abbreviations. For example, rather than typing “University of Pittsburgh” hundreds of times in a policy manual, one might define UP as an entity that meant “University of Pittsburgh.” Entities appeared in the text preceded by an “&” and followed by a “;.” The other purpose of entities was to allow the author to include reserved characters such as “&” and “<” in the text. When XML was rewritten, including entities in the definition was deprecated but those required for special characters were defined globally – “<” for “<” , etc. XML also simplified and systematized the process of writing definitions such that the same program that parsed documents could be used to parse the definition of a document. XML made other powerful improvements such as namespaces that are beyond the scope of this discussion. XML’s definitions of documents are no longer called DTD’s but are called “schema” making a clear parallel to database technology which defines a data table via a schema. XML also introduced restrictions of the type of content an element could have which has made it a desirable vehicle for exchanging data between applications of all types.

For our purposes, the precise language used to define a document schema is not as important as what it accomplishes. A very simple definition of a document might be something like “A “memo” consists of a “heading” followed by a “body.” The heading consists of a “to”, “from”, “date”, and “subject.” The body consists of one or more “para.” We could write this as follows:

memo := (head, body)
head := (to, from, date, subject)
body := (para+)
to := PCDATA
from := PCDATA
date := PCDATA
subject := PCDATA
para := PCDATA

The last five lines use a reserved word “PCDATA” to indicate that these elements will contain no more elements, but only parseable character data. The “,” between the elements in the first two lines indicates that the various elements must appear in the order specified. A “|” could be used to indicate a choice of elements. For example, we could say that a memo would be made up of a head followed by a body or an image as follows:

memo := (head, (body|image))

These operators are called sequence operators. There are also a series of occurrence indicators. The "+" sign shown above means one or more. The "*" means zero or more and a "?" means zero or one. (XML actually now allows a much more complete definition of occurrences. It is possible to write schemas that specify something will occur a minimum of three times and a maximum of 7 times. It is also possible to define restrictions on the content allowed in an element – at an excruciating level of detail.) Using grouping, sequencing, and occurrence in a little more realistic example, we might define something like the following:

MYDOC := (INTRO, CHAP+)
INTRO := (TITLE, SUBTITLE?, AUTHOR, INST?, DATE?)
CHAP := (CTITLE, (PARA|LIST|EXAMPLE)+)
LIST := (ITEM+)
TITLE, SUBTITLE, AUTHOR, INST, DATE := PCDATA
CTITLE, PARA, LIST, EXAMPLE ,ITEM := PCDATA

The example above is still just a toy. There are any number of things wrong with it. For example, what if there are multiple authors? Should we separate the first and last names of the authors? Should we make middle initials optional? Should an institution element, if present, include the main institution and allow for a subordinate unit. Should there be sections in chapters? Should we allow paragraphs in the chapters but outside the sections? Should there be a table of contents? It is not hard to see that defining a good schema for documents is not trivial matter. What is required, what is ordered, what is allowed are all complicated questions, especially if the definition is to work for a large number of different documents. The benefit of working out a good definition is algorithmic processing. It is possible to instruct a program where to look to find the name of the author – something that simply can't be assured in HTML.

Marking up a Document

Once we have a schema, we can begin to markup a document. Beginning with some content:

A Document
Course on E-Business
by: Michael B. Spring, University of Pittsburgh
General Description
The Internet is providing new ways of communicating and of doing business. There are many facets to the developments and many technologies appearing and disappearing in the rush to develop this new area.
The objectives of this course will be to:
Understand the pressures that are moving e-business forward as a new mode of doing business
Appreciate the impact of bit businesses versus atom business, national versus global markets, and customer driven manufacturing on the conduct of business
This provides an overview of where we are

Figure 22: Simple Text Document

To that content, we will add some structure. Three things can be noted about the elements. First, every element has both a begin tag and an end tag. Second, elements do not overlap. That is, content belongs, or is the child of only one element. Third, elements either contain text, or they contain other elements, but not both. In a real schema we would allow some elements to have a mixed content model – both other elements and text. For now, we are trying to keep it simple.

```
<MYDOC>
        <INTRO>
                <TITLE>Course on E-Business</TITLE>
                <AUTHOR>Michael B. Spring</AUTHOR>
                <INST>University of Pittsburgh</INST>
        </INTRO>
        <CHAP>
                <CTITLE>General Description</CTITLE>
                <PARA>The Internet is providing new ways of communicating and of doing business.
                There are many facets to the developments and many technologies appearing and
                disappearing in the rush to develop this new area. The objectives of this course will be to: </PARA>
                <LIST>
                        <ITEM>Understand the pressures that are moving e-business forward as a new mode of
                        doing  business</ITEM>
                        <ITEM>Appreciate the impact of bit businesses versus atom business,
                        national versus global markets, and customer driven manufacturing on the conduct of
                        business</ITEM>
                </LIST>
                <PARA>This provides an overview of where we are:<\PARA>
        </CHAP>
</MYDOC>
```

Figure 23: Structural Markup and Text

This structure can be visualized as a graph beginning at the root (MYDOC) and terminating at the content nodes. The tree for our little example is shown below:

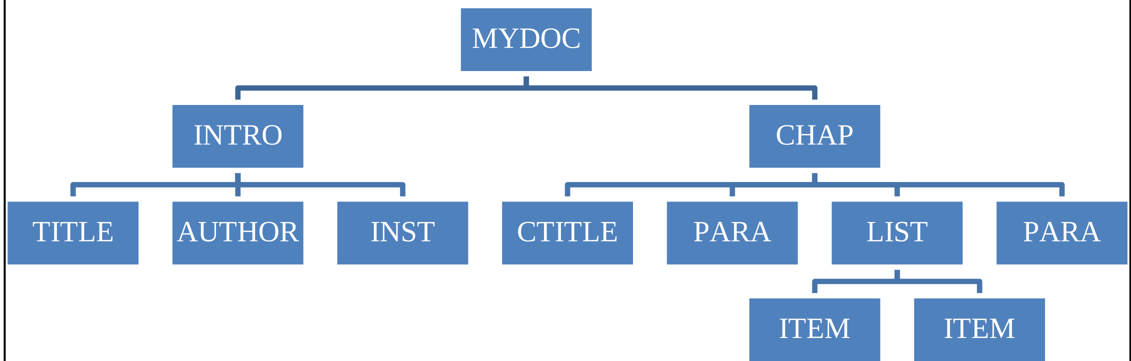


Figure 24: Graphical View of a Document

Processing a Document

So far, the structuring of documents looks to be a lot of work, and indeed it is. Tools are emerging that make the process easier, but there are not many good document definitions and there are fewer tools that make it easy to write. Reflecting back on working with Scribe, even though commands had to be typed, Brian Reid's rather elegant and parsimonious definition of various types of documents made it easy to do all the things that I wanted to do without being distracted by the markup. While better editors are emerging today, we still need better and more disciplined schema definitions.

We have not addressed two other significant parts of the XML standards suite – data types and transformations. Data types were introduced in XML that allow the specification of the kind of content that might exist in an element. In SGML, the content of an element was outside the scope of the standard. In XML it may be specified. This kind of typing can be done within an element such as following:

```
<element name = "pubyear">
        <simpleType>
                <restriction base = “positiveinteger">
                        <minInclusive>1000</minInclusive>
                        <maxInclusive>3000</maxInclusive>
```

```
        </restriction>
    </simpleType>
</element>
```

This defines an element named “pubyear” and restricts it to integers between 1000 and 3000. More complex data types may also be defined. For example, the following describes a 14 digit phone number including the parentheses, space, and hyphen. The characters must be digits.

```
<xsd:simpleType name="TelephoneNumber">
    <xsd:restriction base="xsd:string">
        <xsd:length value=“14"/>
        <xsd:pattern value=“\(\d{3}\) \d{3}-\d{4}"/>
    </xsd:restriction>
</xsd:simpleType>
```

It is this ability to build and use precisely defined data types that has caused the development community to look to XML as a language that can be used to exchange controlled information between systems.

The XML Style Language for Transformation (XSLT) is the other aspect of XML important to mention. In a sense XSLT provides the significant benefits that accrue from the use of a tree structure for the document. The word style in the name of this part of the standard comes from the fact that originally, XSLT was defined as the replacement for one of the SGML family of standards – the Document Style Semantics and Specification Language (DSSSL). Because SGML and then XML were restricted to logically structuring documents, a separate set of conventions were required for the layout of documents. This was to be the job of DSSSL and then later the XSL Formatting Objects language. For all practical purposes, Cascading Style Sheets have emerged as the preferred way to provide Styling of XML documents, but the Style Language for Transformation remains an important part of XML.

XPath provides a declarative language base for XSLT. To understand XPath and XSLT takes some time, but a simple introduction should suffice to give you a sense of the power they provide. If we go back to the tree model shown on the previous page, it is possible to define a “path” to the chapter title. We could write it as follows:

/MYDOC/CHAP/CTITLE

That is, the CTITLE is the child of the CHAP which is in turn the child of MYDOC. This is the most primitive XPath expression. The expressions can be made more expressive in a variety of ways. For example, the “/” used above is shorthand for one of the thirteen axes that define path traversal in a tree. We can find the child of a node, or it parent. We can find the element that precedes or follows and element or we can find the next or preceding sibling. As one example consider the axis that allows you to define all the descendents of a given node:

/MYDOC//ITEM

This expression, note the two "/"'s between mydoc and item will locate all the ITEM nodes that exist in the tree anywhere under the MYDOC node. XPath has functions and predicates as well that make it possible, using the transformation language to turn documents inside out and upside down. As we will discuss in the section of the future of documents, this capability makes it possible to build and use documents in totally new ways. One more example may serve as an indication of that power.

.//footnote/../title

Would select all of the title nodes that were children of all of the nodes that were parents of the footnote nodes that were descendants of the current node. More formally this would be written as:

self::node()/self -descendant::footnote/parent::node()/child::title

Other Parts of the XML Standard

There are many other standards related to XML that have been or are still being developed by the World Wide Web Consortium (W3C). They fall into two categories, core standards and derivative document types. The XML base standard defines the schema, data types, and basic formatting rules. This group also includes provision of the ability to use multiple schema in a single document – something not possible in SGML. In XML the capability falls under the general category of "NameSpaces." The XSL standards define styling and transformation as well as Xpath as we have described above. XLink and XPointer provide extended capabilities to provide hyperlinks between documents. Imagine that the kinds of XPaths we have described above could be used to define a specific fragment of a document. That makes it possible to create a link not just to a document, but to a range or element within it. It is also possible using the XLink specification to create a variety of typed links that allow one anchor in a document to point to multiple places for different reasons. XForms provides a new way to define forms and data to be shared when a processing request is made. These forms share the general structure of separating structure and appearance. Finally, most processing of XML documents is done on individual documents. XQuery provides mechanisms for using XPath across collections of documents.

The other work that is being done with XML at the W3C looks to the development of document types that serve special purposes. For example, the Mathematical Markup Language, (MathML) provides markup for mathematical notation and capturing both its structure and content. The goal of MathML is to enable mathematics to be served, received, and processed on the World Wide Web. The Synchronized Multimedia Integration Language (SMIL) allows authors to write interactive multimedia presentations. It provides a means to describe the temporal behavior of a multimedia presentation, associate hyperlinks with media objects and describe the layout of the presentation on a screen. This same group has moved on to create the Timed Text Markup Language (TTML), a document type that represents timed text media for the purpose of interchange among authoring systems. Timed text is textual information that is intrinsically or extrinsically associated with timing information. Finally, my personal favorite is the markup language for Scalable Vector Graphics (SVG). SVG is a language for describing vector graphics – mathematical descriptions of graphical objects that allow them to be specified in great detail. Most importantly, unlike raster graphics – images, vector graphics can be made larger or smaller without any loss of precision. The language allows for a variety of operators and blending modes that produce stunning graphics.

Digital Documents

As these standards emerge and tools based on them are developed a whole new world of documents becomes possible. The era of structured and typed digital documents has begun. Initially, these new documents will still be bound by older visions of documents. In the next section some of the radical changes possible in the world of documents will be discussed. For now, we will look at the changes that we can begin to incorporate today.

Document Classes

In what follows, it is important to distinguish between documents in several broad classes. One way to classify documents that captures many of the differences that naturally occur is by the content in which the document exists. I have found it useful to classify document s as:

- Personal
- Group
- Organizational
- Enterprise
- Archival

As you might imagine, it is less likely that highly structured documents will be used for documents that are personal. I don't necessarily need all the power of structured documents when I write poetry. On the other hand, when I co-author an article, some level of agreement is required about how we will write. Beyond personal and group documents, organizational documents may be structured in accord with the processing needs of the organization. The continuum continues to enterprise documents – documents that are exchanged between organizations, such as a purchase order. Finally, at the end of the continuum we have archival documents which should be so constructed as to be accessible into the future for access by different people for reasons we may not be able to anticipate today.

These goals, while academically easy to theorize, are still difficult to achieve – old habits die hard. Consider just one example. The vast majority of faculty in higher education who offer a course develop a syllabus for the course. This guides students in knowing what will be lectured on each week, when assignments will be due, what the grading policies are, etc. For more than twenty-five years, I have been preparing syllabi for my courses. Each time a course is offered, a new syllabus is written based on the old one. Given my area of research, you might guess that the documents are highly structured and highly detailed. Faculty have been writing syllabi for decades. Each term, I ask my students to write a schema that would provide a viable structure for syllabi for all faculty. The task is extremely difficult and many of my colleagues would suggest that such detailed descriptions border on an infringement of academic freedom. Far from it, I believe such detailed syllabi should be the hallmark of academic responsibility. Regardless, there has been no great rush to adopt a standard schema for course syllabi. (Indeed, I note that despite great pressure from the Registrar at my own institution, there is probably still too much variance in what satisfies the requirements of a course description!) That said, we will gradually see the emergence of standard document forms in the organizational, enterprise, and archival contexts. Indeed, those standard forms that have appeared have been in these environments.

Types of Content

In the typewriter world, documents were predominantly text based. A graphic or an image might be included, but they were, like the text static. Today, digital documents are increasingly dynamic. While it is not done extensively – note my own laziness in this book, a book can contain spreadsheets and not just printed tables. This would make it possible to check the logic behind a table, or to graph the data that exists in a table. Similarly, a graphic can now become an animation. One of the times a white board is still superior to a PowerPoint presentation is when it is useful to have the student see a calculation or a flow chart develop. Some teachers have developed presentations that look to capture this flow. But today, the cost of developing these dynamic presentations is still high. With time and the development of better tools, documents will become more dynamic and malleable. Indeed one of the most important developments of the last fifty years has been the ability to include objects of various types in a digital document.

What Digital-Documents Can Do

When one talks of digital documents, it is not unfair to think about e-books. Digital documents will ultimately be much more than e-books, but that is surely not a bad place to begin. E-books already do many of the simple things that digital documents will be able to do. The system can keep track of where you left off reading. I can make annotations on the document, and bookmark pages or paragraphs that are of interest. I can share what I have found with others and I can ask the system to summarize what I have written. Of course, search for words is easy, both within a document and across documents. Highlighting is in some ways better and in some ways still lacking. The ability to highlight a paragraph in yellow or light blue is easier than ever, and neater. The ability to use "*" of infinitely varying sizes is not quite so easy. The ability to look up a word is a simple task, and with a network connection, which we will assume for digital documents of the future, e-books can easily link to online reference works such as Wikipedia. Translating a term or phrase to some other language is a capability that has already been demonstrated by Google. Text to speech technology makes it possible for my book to read itself to me, and speech to text will make translated audio annotations the norm. E-books also allow cross references in the form envisioned for Vanevar Bush's Memex, but screen territory remains an obstacle to easy cross reference comparisons.

While finding where a word occurs, or defining it is pretty straight forward, having a digital document find other "relevant" material requires significantly more effort. Good models are emerging for this kind of reference through algorithms such as those employed by Netflix, but identifying documents that are semantically related is still more complicated than finding documents with similar word patterns.

The next step in digital documents is giving them more knowledge about the reader. Knowing who I am opens up a variety of new possibilities for digital documents, as well as opening up new concerns about privacy. For example, a digital document, knowing what I know might suggest remedial or advanced treatments of a given topic. If who I am is known, and all other readers are also known, the document might suggest a conversation with someone like me who is reading the document at the same time. We might be given summaries produced by other readers. Digital documents in the network might be able to pass a question about the book onto someone who could answer it. It might keep track of questions and answers so it can answer a question by itself at some point in time. It might be able to show me how to carry out some process or calculation. It could expose the information on which some graphic, or statistics are based. Finally, it might reorganize itself based on reader interests or capabilities. Introductory material could be put in endnotes, or advanced material could be exposed.

Documents of the Future

A likely future for documents, and other media types, is increased dynamism. Not only will documents become digital and structured, but they will become active. By this we mean to suggest that rather than users having to locate documents, an appropriate infrastructure could allow for the development of dynamic documents that could locate humans who might be interested in them. As just one simple example, imagine a web based document where scroll bar movement was recorded and fed back to the information store responsible for the document. Imagine that 30% of the viewers of the document scrolled back multiple times from a page where a concept was mentioned to the location in the document where the concept was defined. Further imagine that the information store noted than 50% of those people used an online source to get more information about the concept. It is possible that this activity might infer the need for more explanation of the concept and different placement of that information. Ultimately, the document might present itself such as to minimize the time I need to read it and maximize the knowledge transferred. It would be long and descriptive for some people and short and authoritative for others. Dynamic documents might know a lot of things and be able to take actions based on what they know. Documents could keep track of who has read or is reading the document, how much time is spent in various sections. Documents could form discussion groups.

More and more documents will be multiply purposed based on a canonical form. Without getting into the technical details of how it is done, it is possible with XML and XSLT to derive one document from another. This means it is possible to create a single document that can be transformed into many different forms. For example, we might imagine a document that describes a system from five or six perspectives – reference, user guide, technical specs, etc. The transformation language would create the needed documents from the single canonical form. Staff would not need to maintain a dozen forms but could make updates to one master document. Similar technology could allow custom catalogs for men and women, etc.

Producing Documents

From an author’s perspective, digital documents might help an author build the basic structure for the book, or find related materials to be included as references. As I am writing this book, I am constantly moving treatments from one chapter to another and placing notes in a "too be considered." It would be useful to have a digital assistant do that for me. Over the last thirty years I have accumulated more than 20 Gigabytes of documents, notes, and resources that are contained in 127,000 files stored on my various computers. It would be useful to have a real time digital assistant look up the lecture on XML and show me the samples I carefully worked out such that I could easily include them here as appropriate.

A published book could report back to the author about how the book might be revised based on reader usage and questions. It might even be so bold as to read the news and suggest new sections to be added as what we know about a topic area changes. The environment should be able to find the data needed before it is asked for it. If I refer to the Negroponte Flop in a paper, especially given the fact that I have read and annotated the book "Being Digital", my system should find the passage and correctly format the reference. Better yet, the system should be able to infer whether I am working on a formal or informal reference and interrupt me as appropriate. (Interestingly, research groups at Microsoft have been working for years on building assistants that understand exactly this –i.e. the value of information over time versus the cost of an interruption.[43])

Agents and Tools

As has already been discussed, rudimentary agents are already assisting us in document processing, but we need more than spell and grammar checkers in line. Research assistants will be used to check out online support for what we are writing, and find the specific information in an article that supports what we are trying to say. All documents should have metadata that allows reference material – the bane of my existence – to be correctly extracted and formatted for the particular publication style. The current state of document editors supporting documents constructed from a schema are overly controlling. Document construction agents will emerge that will wrap around the author and gently move them toward insuring that all the needed parts get put together as they need to be to meet the schema requirements.

Collaborative support agents will become more sophisticated and will help teams work together on activities. They will keep track of what we do and let the other team members know what we are doing. Translation and integration of mail and meeting minutes will be seamless

[43] E. Horvitz, Eric et.al. (2004) BusyBody: Creating and Fielding Personalized Models of the Cost of Interruption, Proceedings of CSCW, Conference on Computer Supported Cooperative Work, ACM Press, November 2004.

and helpful. Running and scheduling meetings and communication will simply be taken care of. Version control and document checkout and check-in will avoid collisions. Editorial agents will observe an author and provide editorial advise or assistance. They will be tunable so as to be minimally intrusive or active in monitoring comments and identifying problem areas or suggesting places for revisions.

Finally, a whole set of new tools will be developed unlike anything we have today. One thing that makes live presentations in a classroom better than canned presentations is the availability of a whiteboard where we can talk through some process or calculation. The drawing and lecture are intimately tied together. When we convert it to a formal symbolic form, we can achieve greater precision, but we lose the intimate balance in the presentation. Tools will emerge that will give us both. When teaching in a classroom, it is easily possible to capture the lecture and the slide presentation. I have considered taping every lecture that I give when I have given the lecture more than three of four times. I don't yet because the "cost" of video editing is still too high. Inserting things and modifying part of the presentation seamlessly still requires too much work, but this will change as well.

Summary

We all know what a document is, but when it comes to providing a clear unambiguous definition, it becomes much more difficult. Documents are containers for information. In 1990, in the predecessor to this book, I proposed the following definition.

> *A document is an identifiable entity having some durable form, produced by a person or persons toward the goal of communication; it may take a number of forms, but must have at least one symbolic manifestation that is used to store or communicate information between people. It is a cohesive entity formed of subcomponents in logical, layout, and content form.*[44]

Historically, an article, report or book consisted predominantly of text and was written by a single author. Things are no longer so simple. Today, documents include text, graphics, images, tables, equations and other objects. They may be cohesive, a letter or a report, or composite, a medical record. They may be authored by individuals, groups, or

[44] Spring, Michael, (1990) Electronic Printing and Publishing: The Document Processing Revolution. New York: Marcel Dekker, p8.

organizations. They may have a limited life span – a notice of a meeting or be archival – a birth certificate. An updated definition for documents might be the following:

> *A document is an artifact created toward the goal of communication that has at least one symbolic manifestation and that may be composed of both static and dynamic components*

If we want to process documents algorithmically, it is important that we have precise definitions. One of the tasks for computer scientists over the last two decades has been trying to develop a formal definition of a document. This chapter has described the movement toward a more technical definition of a document. A document is:

- A logically structured set of elements that take the form of a tree, or more formally a directed acyclic graph (DAG).
- The permitted nodes are defined via a document type definition, or schema.
- The nodes of the graph are named and may contain content or other nodes.
- The nodes may be described via attributes which provide information about the document
- The leaf nodes of the tree contain the content of the document.

Chapter VII: The Revolution

Introduction

Let me start with a simple proposition – "the printing press led to a radical change in the cost of preserving and disseminating information." Some have selected the printing press as the most significant invention of the second millennium.[45] In large part this is because of the impact the printing press had on the dissemination of information. Elizabeth Eisenstein's "The Printing Press as Agent of Change"[46] provides a wonderful account of the upheaval. The printer scholars had an impact on everything from the Protestant Reformation to the Scientific Revolution.

Just as the printing press led to a revolution in information handling, computers and networks have introduced another revolution. The nature of this revolution is worth a little exploration. The last two decades of the 20th century saw several changes in printing and publishing. First, in the mid 1990's, the web challenged the role of the publisher. It became possible for every individual to be their own publisher. There are pros and cons to this development, but the fact is that the publishing process was challenged and to a great degree disintermediated. Second, publishing moved from a push process to a pull process. It was no longer a matter of marketing so much as interest. These are complicated changes and not quite so absolute as stated here, but there is none-the-less a dramatic change in the historically absolute

[45] Many of the news stories were based on work the auspices of www.edge.org. The discussion may be construed as offering a slight edge to the printing press, but the correspondents in the discussion surely offered many different alternatives. There was little agreement about what constituted an invention and one might suggest different answers based on whether one was looking at ideas and concepts, physical objects, or discoveries. In the arena of ideas and concepts, zero, calculus, and the scientific method all made great sense. From the point of view of discoveries, the germ, DNA, the atom, and electricity are all hard to beat. But if we restrict ourselves to physical devices, television, the computer, the steam engine, the electric dynamo, the light bulb, and the printing press come high on the list. I would concur that the printing press has a slight edge.

[46] Eisenstein, Elizabeth (1979) The printing press as an agent of change: communications and cultural transformations in early-modern Europe. Cambridge University Press.

gatekeeper role played by publishers. Related to this change from push to pull has been a shift in the cost structure for dissemination, with a greater share of the cost falling to the consumer – the fax and printer paper, connect time, the computer and printer, etc. Third, there has been a significant decrease in the cost of bits of information. At the far end, it is possible to suggest that under certain conditions, the cost can be projected as approaching zero. For paper-based dissemination, the cost is the cost of the consumable, let's say a penny/page. With electronic dissemination, the cost is the cost of the electricity to power the computer and monitor – far less than a penny per page. There are actually more costs associated with installing and maintaining the infrastructure, which will be considered later.

The digital information revolution shows some indication of being more fundamental than a revolution in printing and publishing, i.e. a revolution in how we communicate – a new era of immediacy. Before looking at the changes in how we communicate, we begin with the revolution in reprographics.

The Reprographics Revolutions

For several thousand years, written communications were unique creations. Surely important works were copied. The history of the Library of Alexandria mentions making copies of scrolls for repository purposes as early as 300 BCE. But copying is not reprographics. The earliest examples of reprographics date back to around 3000 BCE with seals used to validate the authenticity of official communications via clay tablets. It seems that the earliest examples of printing, beyond the use of seals, was to reproduce patterns on cloth. With time, "woodblock" printing grew from printing images on cloth to printing text on paper. Examples have been documented in India and China as early as 800 and 1300 CE. Composite blocks, i.e. moveable type, appeared in China around 1040 CE and in Korea before 1300 CE. While all of these developments were important, the explosion in reproducing documents did not catch on until Gutenberg joined new methods of composing the moveable type. Beginning with Gutenberg, the development of reprographics may be broken into four ages.

1. Mechanical Reprographics
2. Optical Reprographics
3. Digital Reprographics
4. Ad Hoc Reprographics

These terms characterize four significant developments in the reprographic process. Around 1450, developments by Gutenberg and others signaled the beginning of the era of mechanical reproduction using modular reusable masters. Mechanical reprographics substituted moveable type and mechanical reproduction for human efforts in reproductions. The printing press replaced the scribe and wood block printing. Mechanical reprographics encompassed a number of different forms. These included a variety of techniques for plate construction as well as a number of approaches that used engraved plates and approaches to offset printing where the ink was “offset” to a rubber roller before being transferred to the paper. This period also saw the development of the rotary press which allowed continuous printing on rolls of paper – rather than sheets.

Optical reprographics represents two distinct revolutions. The first was photolithography, the use of photographic technology to make printing plates. The lithographic process is older than photolithography, and is basically a process invented in 1796 that allows a smooth surface, initially limestone, to be prepared in such a way that the surface is receptive or non-receptive to ink. When the lithographic process was combined with the photographic process, it became possible to use the technique to produce a printing plate from an image. While this would initially be devoted to producing plates for images, ultimately it would lead to phototypesetting. This development would provide a cost effective alternative to mechanical typesetting causing a significant change in the printing industry. The changes took some time, with lithography in manually produced form emerging around 1800, various methods for creating lithographic plates using photography emerged during the 1800s. During the 1900’s several improvements were made that allowed the development of photographically produced plates. Notable among these was the use of metal plates and offset printing. (Offset printing preserved the metal plates from the wear and tear of paper rubbing against the plate by offsetting the ink to a rubber blanket which then transferred it to the paper.

The second part of the optical reprographics revolution coupled photographic technology with photosensitive dielectric materials. Dielectric materials are a class of materials that will hold a localized electrical charge. Put another way, once a charge is applied to the material, it will not move. In 1938, Chester Carlson used a photosensitive dielectric, amorphous selenium, that held the localized negative charge, and dissipated it when exposed to light – where the material was not exposed to light the charge stayed. If we imagine, as shown in Figure 26, the selenium forming a drum that is rotating, and a material of opposite charge (positive) is then placed near the selenium, it will be attracted to those places that had not been exposed to light. (If you remember rubbing a comb against your hair to build up a static charge which you could then use to pick up pieces of paper, you can imagine that finely ground black plastic as the material (toner) placed near

Figure 25: First Copy by Carlson

the selenium.) The selenium drum is then rotated over a sheet of paper that has a very high negative charge behind it causing the finely ground plastic to be dragged to the paper. The paper is then passed between rollers that are heated, melting and pressing the toner into the paper. The result is photo duplication commercialized by Xerox. This made it possible to take any object that could be placed on the glass of the photocopier and transfer its image to a reusable master plate that could be used for the printing process. The revolution here was dramatic in a number of ways. From a technical perspective, the master printing plate was reusable. Setup cost was virtually eliminated. From a social perspective, virtually anyone could become a printer.

The third revolution, digital reprographics includes the fax machine, laser printers, and most recently digital copiers. The fax machine is interesting in that its roots go back to the 1850s. I selected Hummel's Telediagraph which was operational in 1898 as the first "modern" fax like device. A legitimate case could be made for earlier developments by Caselli's pantélégraphe or Bidwell's scanning phototelegraph both of which predated Hummel. Early fax machines were used primarily for photo's, and used telegraph lines or radio waves for transmission. In 1966, Xerox introduced the Magnafax Telecopier which was designed to operate over phone lines. Through the 1970's that telephone based facsimile transmission grew in popularity. Meanwhile, in 1969, at the Webster Research Center of the Xerox corporation, a researcher named Gary Starkweather was working on the first laser printer. Not well suited to the

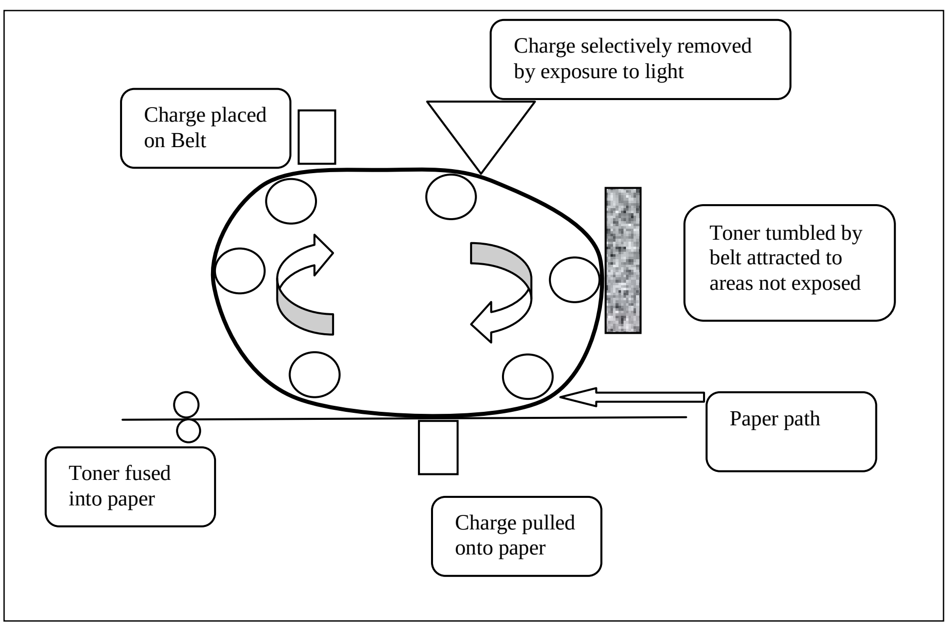


Figure 26: Photocopy Process

research environment at Webster, Starkweather managed to get to Xerox PARC where he had a functional laser printer operating by 1971. With the introduction of digital technology, it was no longer necessary that the original have a physical form. If it did have a physical form, the revolution made it possible for the original and the copy to be in different locations.

I call the most recent reprographic revolution "ad hoc reprographics." The term is meant to capture the form of reprographics where both the master and the copy are reusable. It is characterized by the "printing" of an electronic master to an all points addressable screen — a computer or TV screen. Ad hoc reprographics is fundamentally different than the earlier stages. No physical resources are consumed. The "document" display is immediately reusable. It is possible that the master and the copy will never exist in a physical form.

The economics of these four eras of reprographics can be summarized by looking at the costs associated with each era. From 1400-1800, the mechanical reproduction, there were two costs involved in making a copy. There was a cost (Y) to make the master and then there was a cost (X) for each copy. Prior to 1453, block printing techniques required an expensive unique master, normally made of wood which had to be carved by hand and wore out quickly. Gutenberg, and moveable type technology made a replicable component based master possible that was cheaper, and being made of metal, more durable. The cost per copy (X) stayed the same as before, but the cost (Y) to produce the master lowered. Over the years, developments in printing presses worked to reduce costs, such as rotary presses to reduce the cost per copy and offset presses to reduce plate wear.

With the introduction of photo-optical processes, the cost (Y) of making the master was again reduced. While lithography was the base technology that made this happen, it wasn't until the emergence of photolithography that hot metal type masters were seriously challenged. As the 19th century technologies were brought together in the 1950's, a secondary photo-optical process was emerging in the form of xerography. The xerographic process was important in that the master became reusable – the setup cost (Y) disappeared.

Since 1950, digital reprographics have continued to reduce costs. There is "no" set up cost and the cost(X) per copy is distributed. As for the Fax machine, it was the emergence of a network accessible to all users – the phone and the development of compression technology that advanced the technology. The laser printer was even more radical in that for the first time it was not necessary to have a physical object that would make use of the reusable master. The final reprographic development appears when we provide screens that are of sufficient resolution that they can serve as a reusable substrate for the copies. Modern PC screen and most recently e-ink based screen present such a substrate.

It is interesting to note that even with ad hoc reprographics there is pressure to maintain the traditional printing, publishing and distribution roles. There are many reasons for this. While publishers don't want to give up their traditional sources of revenue, it is equally clear that they play what many consider important roles in helping the author clarify, present, and market their message. Another factor has to do with aesthetics. As Donald Knuth noted when he looked at a computer typeset book next to one set with hot metal type, there was a significant difference. Others have noted this as well. While computer typesetting is dramatically better today, in part thanks to Knuth, there is still a noticeable difference in the products of the various eras. Another reason why traditional models remain is simply the inertia of traditional organizations and habits.

While the revolution on reprographics is compelling, it may be the case that a more radical revolution is afoot.

A Revolution in Communications?

While the changes in the cost of reproducing and disseminating information are important, there seems to be something more fundamental and significant happening. A number of authors have suggested that radio and television, which few would deny have had a significant impact on society, have had a profound impact on communications. This can lead one to think about the current revolution as a revolution not in methods of reprographics but in methods of communication. We know that things changed radically when humans developed language. It likely did not happen all at once, or in one place, but it is an interesting exercise to try to imagine two tribes, one grunting and gesturing and the other using words to explain what needed to be done. How much were the development of farming and herding a secondary effect of language development? There is no doubt that the amount and quality of aggregated knowledge was greatly accelerated by the development of writing. In the last two hundred years, some additional developments have taken place. One is the ability to record an experience for replay later. Another is the ability to communicate instantaneously. We begin this section with a review of the oral and literary traditions and then look more intensely at what is going on in our lives today.

The Oral and Literary Traditions

There is evidence that "humans" have been around for a million years or so.[47] It is difficult to pinpoint when spoken language developed as a critical means of communication. Ong holds that social interaction has been occurring for the last 30,000 to 50,000 years.[48] Although it is impossible to date the origin of spoken language, it was clearly many millennia before written language. Scholars refer to this period when spoken language was the dominant form of communication as the oral tradition. Imagine communication prior to spoken language. There is not much that easily comes to mind, but there must have been something – the pantomime of the great hunt, the defeat of the great enemy, etc. These communications must have been subject to significant misinterpretation. One might speculate that the creation stories all have some common origin. They seem to be bound to physical objects and dramatic events that could be described without language. They would involve physical universals — the sun and the stars, catastrophic events such as fires and floods, etc.

The oral tradition, from an information theoretic perspective allowed information to be codified by means of language and memory and passed on from generation to generation. However, the oral tradition was subject to information loss in the reproduction process, and the capacity of human memory was a limiting factor in the amount of information that could be passed on. Further, the transfer required that sender and receiver be collocated in time and space. The literary tradition eliminated the need for collocation, vastly expanded the amount of information, and made a significant improvement in transmission errors, although it left coding and decoding errors.

The oldest deciphered written documents are about 6000 years old.[49] Scholars refer to this form of communication as the literary tradition. The literary revolution occurred when symbols where used first to represent objects – pictographic writing and then to represent spoken language – phonetic writing. After 20,000-30,000 years of the oral tradition, the development of writing must have caused quite a stir. Imagine the court poets arguing over the value of these new scribes. I envision a test where the powers in the court, based on the oral tradition, ridicule the fact that a message written by one scribe can't be read by another. It would not be too

[47] Fromkin, Victoria and Rodman, Robert (1993) **An Introduction to Language**. Fort Worth, Texas: Harcourt Brace College Publishes, p. 22.
[48] Ong, Walter J. (1982) **Orality and Literacy**. London: Methuen & Co., p.2.
[49] Fromkin, Victoria and Rodman, Robert (1993) **An Introduction to Language**. Fort Worth, Texas: Harcourt Brace College Publishes, p. 22.

different from the scorn we heap on computers that can't communicate. The message in spoken form was easily understood by all. Memory in this frame must have been viewed as a far superior mechanism for communication. The human store could be queried, it was easily understood, etc. It is not hard to imagine the highly valued court poets claiming that anything recorded in this new way would surely be lost because even two experts could not agree on how to record a spoken word. Any individual — king to slave — could speak and listen, but only specially trained technologists could read and write. This picture has a modern counterpart. Technological advances never come automatically.

Many years after the invention and wide dissemination and acceptance of writing, Plato has Socrates commenting on the technology:

> *This invention of yours will produce forgetfulness in the minds of those who learn to use it, by causing them to neglect their memory, in as much as, from their confidence in writing, they will recollect by the external aid of foreign symbols, and not by the internal use of their own faculties.*[50]

Many people find this quote attributed to Socrates humorous. It was so long ago, and so ludicrous given the current state of wisdom. However, even to this day, there are places and situations in which the oral tradition prevails, for example, the notion of eye witness testimony versus affidavits. People had to be persuaded that written documentation was a reliable reflection of concrete, observable events. To the modern mind, the evanescence of the spoken word seems more plastic, quixotic, and undependable than the printed word. To members of a highly oral culture, however, the spoken word was connected to the incontrovertible realities of bodily experience, while the written word was a thin, substance less scratching whose two-dimensionality seemed highly arbitrary.[51]

At the same time the written word might reduce memory skills, it is a powerful technology to supplement and improve the communication process. The revolutionary change in communication occurred with the virtually unlimited storage capacity and permanence represented by writing. The literary tradition was accelerated with the development of mass produced books in the fifteenth century. While mass production printing was not a new form of communication the impacts of opening up the communications channels was dramatic.

[50] Plato, Phaedrus, on Socrates, sect. 275
[51] Zuboff, Shoshana, (1988) In the Age of the Smart Machine: the Future of Work and Power. New York: Basic Books, p. 77

Telepresence

The oral and literary traditions evolved over the years each assuming its most appropriate roles. In the nineteenth century, the invention of the telegraph signaled the beginning of a new form of communication. We will peg the beginning of this new era with the transmission by Morse of the message "What hath God wrought" on May 24th 1844 from Baltimore to Washington to open the telegraph line that had been commissioned by Congress. The telephone "quickly" followed beginning in 1878 and was making significant penetration by 1900. In 1900, broadcast radio was added to the mix evolving from wireless telegraphy. Experimental stations appeared beginning in 1909, and licensed stations appeared in 1920. For the first time, virtually instantaneous communications was possible over distances. Ong suggests that the technology was encouraging a second orality.[52] Some have suggested that this new orality would cause a decline in the literary tradition. The decline in the number of general circulation newspapers was seen as one indication of a reversion from written to spoken sources for current events information.

Evolving from a series of electromechanical devices, the first electronic scanning television was developed in 1927. By the mid 1930's television had emerged as a stable technology, broadcasting events at the 1936 Olympics in Germany. Immediate, global television broadcast of events ranging from the Olympic games to the Persian Gulf war represent something of a return to the pre-oral tradition. They represent a form of telepresence. Within a few decades, after a series of preliminary forms, magnetic tape evolved as a mechanism to record audio (1930's) and video (1950's) broadcasts. This recording is in the same form as the original. It is no longer necessary to retell the story, it can be replayed exactly as it happened. This can be very helpful in the analysis of transient events. Whether it is Armstrong stepping onto the moon, the Kiwi's winning the America's Cup, or a human or natural disaster occurring somewhere around the world. We now have the ability to be present to the event — neither the oral or written tradition has to be relied upon. In addition, because signals can be stored, the events can be replayed and analyzed repeatedly. There are limits to what is recorded. For film or video, the camera angles and lighting must be correct or there may be important information that is not on the tape.

Over the years, audio, video, and photographic technologies continued to evolve. In the 1960's and 70's, based on telecommunications

[52] Ong, Walter J. (1982) **Orality and Literacy**. London: Methuen & Co., p. 136.

technologies, a digital form of audio recording took shape. During the same period photographic processes moved to digital form – based on the needs of spy satellites. Electronic cameras, initially storing an analog recording, moved to a digital form that became available commercially in 1990. Digital video recording appeared about the same time although true digital video cameras would not be available for widespread use for another decade. Over the last century, electronic transmission and recording of events in the real world have developed a separable communication tradition that in a sense obviates the need for a human recorder and reteller of the event. Within the limits of the capture, transmission and replay, these technologies represent a form of telepresence. When recording is added, that telepresence is allowed to be asynchronous. When all of the recordings become digital, it is possible to merge them with the literary tradition.

Transliteracy

It could be argued that early computer mediated communication is a return to a literary tradition. The proliferation of electronic mail contrasts with a decline in letter writing and an increase in telephone usage. Hypertext is a return to knowledge expressed in a written form. The majority of World Wide Web resources were initially text based. This material was highly dynamic. Documents are being changed frequently, new documents are added and old ones deleted. (Remember when lots of links were broken?) This period has largely passed as we learn the importance of link permanence.

At the same time, the documents stored on the Web are different than those of the earlier literary tradition. On the Web, the generation and management of links approaches the significance of the text itself. This emphasis on linking is dramatically different than the citations and references that were the equivalent in the literary tradition of the second millennium. Not only do we construct a communication, but a structure that provides multiple ways of navigating it.

Beyond the linking and structuring of the communication, modern documents may be easily composed to include video, audio, graphic, and image perspectives. There is an increasing use of audio and video documents to supplement text based documents. This process is envisioned by Ong as well as Marshall McLuhan.[53] More recent researchers have called this development transliteracy.[54] So, what is happening today might

[53] McLuhan, Marshall (1964). **Understanding media: The extensions of man**. London: Routledge.
[54] Thomas, Sue et al, (2007) "Transliteracy: Crossing Divides," First Monday 12 (Dec. 2007). Retrieved May 2011 from the firstmonday website: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2060/1908

be characterized by three things. First, the communication may now encompass both symbolic and oral communications. Second, it may provide a kind of telepresence in which the impact of the sender on the experience being shared is minimized. Third, the communications may be customized by providing multiple paths for the receiver to use in navigating the communication. Beyond these major features, the author can now weave processing instructions into the communication. These enable the receiver to understand how the author developed the communication or that allow the receiver to continue to modify and change the document as it is accessed. I would suggest that an appropriate name for this new tradition might be immediacy.

The Era of Immediacy

The proposition here is that the development of communication based on digital information may be as significant as the development of writing. This is a rather audacious claim, but I believe we will look back in a few hundred years and say that the last decade of the twentieth century saw the rise of a new communication tradition – immediacy. It is based on a much longer period of development which can be traced back to the development of the telegraph in 1850. So what then is immediacy? Keep in mind, the word "immediacy" is suggested merely as a placeholder. Time will produce a better or more appropriate term, but the concepts that might be associated with "immediacy" are clear, especially those that provide a contrast to prior traditions. They include the following:

First, in the long tradition of communication via the spoken and written word, the communication of information is via an intermediate party. The information gained is not immediate. The information passed from generation to generation via the oral tradition was rich and structured by the orator or storyteller.[55] The tradition and the techniques are fascinating and well beyond the scope of what can be covered here. For our purposes, the key is that a story that was passed on was about events as interpreted by the storyteller. The literary tradition has the similar characteristic – it is an interpretation of events presented in a symbolic form.

Contrast these presentations of information with the broadcasts of the Hindenburg disaster, the Kennedy inauguration speech, the O.J. Simpson car chase, etc. All of these events are presented without

[55] One of the many fascinating accounts is provided in: Ong, Walter (1981) **The Presence of the Word.** Minneapolis: University of Minnesota Press.,

intermediation – they are immediate. Of course, it may be argued that there are interpretations provided by the decision about where to place the video camera or the microphones. In the case of something like television news, there are questions about how the video is edited or the context in which it is put. None-the-less, there is a qualitative difference with which we may be exposed to information rich events. Take the growing presence of webcams on the Internet for viewing public places or traffic flows as other examples of immediacy in the communication of information. I don't need to be told, or to read, that it is raining in Pittsburgh, I can simply look at the screen and see the rain. Indeed, it is possible to learn that it was raining yesterday or now by connecting the live webcam or its archives from anywhere in the world. This aspect of immediacy relates to my presence to the event. As one final example, consider that major events are now being captured by multiple cameras. In 2011, fifty cameras, many of them mobile, were deployed to capture the action at Super Bowl XLV. Some sporting events are now simultaneously broadcasting multiple video feeds allowing the remote viewer to decide how they want to view the event. The recent move to 3-D television is viewed by many as the gateway to holographic broadcasts that would allow us the kind of experience immersion envisioned by the holodeck of Star Trek fame.

The second aspect of immediacy relates to the speed with which the information may be disseminated. We now have information floats of seconds where historically the lag between the event and the information about the event was in terms of days or weeks. War and space coverage are examples of this. In the 2003 Iraq war, viewers were able to get a picture of an advancing tank column as it occurred. Indeed, two of the events burned into the memory of 50 year olds in the year 2000 are the funeral of JFK and the landing of a man on the moon. Both of these events received wide and immediate coverage. This sense of immediacy refers to the temporal nature of the communication.

A third aspect of immediacy has to do with the immediacy of a vast information store to the creator of a message. Many of us are familiar with the process of dragging and dropping information from one place to another on our electronic desktop. Many of the sections of this book have been created by dragging and dropping parts of lecture notes and slide presentations created over the last two decades of teaching and researching in this area. While I grow increasingly concerned about the loss of information created on very early systems or using now defunct information formats, this is, I believe, a temporary phenomenon. With time, we will have increased more immediate access to the stuff of information and knowledge creation so as to more effectively grab and convert it into messages. The day is not far away when lectures might be captured as a matter of course. A little, but not much further away, is the time when I will be able to get to the end of a good lecture and say "That was a great instantiation of the ideas I meant to convey. Convert it to written form, insert my diagrams, show the steps I suggested for processes, and animate the two critical development sequences." This is immediacy in information creation.

A fourth aspect of immediacy has to do with the receiver's access to the message as an evolutionary whole. This concept is a little harder to describe than the others because, while important, it is not one that we have seen in practice very much as yet. Historians have a fascination with drafts of important speeches. An edited copy of an inaugural speech or a speech like the Gettysburg address provides an opportunity to try to interpret the thought process that underlies the communication. Historically, the process of depicting the evolution of a communication has been very hard. For years now, we have had the capability of easily capturing the version tree of a document in process. With time, it will be more common to have access to the complete record of the development of a message. When that occurs, recipients of a communication over space and time will have the ability to "see" a communication evolve in the mind of the sender in ways that we can barely imagine today. The ultimate implication of this aspect of immediacy is likely to be living documents that capture the creator's efforts and allow the receiver to query not only the document but the thought process behind the development of the document.

A fifth aspect of immediacy has to do with all of the digital aspects of the form. The communication can be repaired on the fly. Errors normal in the transmission can be detected and corrected on the fly. The communication has developed a degree of immunity to the noise in the communications channel. Thus, we can now hear a pin drop in a conversation with an individual on the other side of the world. This in not because there is no noise in the communication channel, but because the information in the message can have additional data added that provides a mechanism for correcting the impact of noise. This same digital quality allows the message to be replicated in its bit form at a fraction of the cost of traditional replication. The message can also be encrypted insuring an appropriate level of privacy and security in the communication.

Summary and Postscript

Over the last hundred and fifty years, we have been developing technologies that are creating the potential for changes in how we communicate. At heart, this is made possible by the development of the ability to represent information in digital form. The core of the revolution is characterized by the immediacy of the communication:

- Immediacy in terms of the ability of the receiver to engage a data stream directly.
- Immediacy in terms of the temporal lag.
- Immediacy in terms of the sender's ability to easily access their personal information for purposes of constructing a message.
- Immediacy in terms of the receiver's ability to investigate the message in its inner details.
- Immediacy in terms of the quality of the communication allowed by the digital nature of the message.

As a postscript, one must always be cognizant of the fact that the impact of technologies can be greatly under- or over-estimated. Further, any truly significant paradigm shift occurs in multiple waves. For example, the industrial revolution that was initiated by harnessing water power did not reach full stride until power from the steam engine was applied. Additional waves occurred with the development of the gasoline engine and finally electricity, both transmitted and stored. It is not clear what we are experiencing today. Is this the final wave of a revolution that began in 1850, or is the development of digital technology in the 1970's the steam engine wave with everything that went before the rather weak water power wave? Are there still significant waves equivalent to the development of the gasoline engine and the electric motor yet to come?

Chapter VIII: Developments to Watch

Introduction

It has been said that the short term impact of a new technology is often overestimated and the long term impact is underestimated. This chapter looks first at some of the developments that are likely to emerge in the coming years. Once grasped, the shifts are easy to see and begin to make sense. However, I find that many people find the ideas presented here hard to comprehend initially. The chapter begins with developments that are more conceptual than technical. A few of these ideas will be explained in more specific contexts in the next three chapters – e.g. the shift from atoms to bits will be discussed in greater detail as it impacts business decisions. In the latter part of the chapter, more technical developments are discussed. The developments selected – secure systems, better interfaces – involve rather safe predictions, as they are already well underway. There will no doubt be other developments that are difficult to predict because they will be based on some simple idea that goes viral.

Atoms versus bits

One of the advantages of digital information is the ease with which it is manipulated and transported. When the underlying business is a bit business the advantages of bits over atoms are so significant that bit versions will supplant the atom based versions. This transition has already occurred for the music industry where the market for CD's is being replaced by downloaded music. It has also occurred in the banking industry where tellers began to be replaced by ATM's and ATM's are now being replaced by online banking. E-books are beginning to take hold, and while it may take a long time for all print books to disappear, that time will come. Physical libraries are being replaced by digital libraries. Maps are being replaced by GPS systems. Yellow pages are being replaced by online business directories. Catalogs and catalog sales are being replaced by online stores. A dramatic impact of the atoms to bits story will be expanded in Chapter XI on E-Business where we will talk about the use of information to replace inventory.

An interesting game involves guessing what bit forms will be next to replace their atom forms. Banking is easy. Stock trading is close behind. The advertizing industry has already begun the transition to bits, most notably through the business model provided by Google. An interesting possibility has been emerging recently through online video phone calls. When my sons graduated from college and graduate school recently we had a party and family reunion. With my family spread across the country, it was not surprising that someone couldn't make it. So we set up a computer on the patio with a camera and my brother was able to partake in conversations and the experience via a video call. Recently my brothers and sisters all agreed to install cameras to allow for conference calls. As the cost of transportation increases it may be that some family gatherings will occur not by transporting atoms, but by moving bits.

We will still buy clothes, food, appliances, etc., things where the atoms cannot be replaced by bits, but early forms of replicators are already emerging. At the current time, they are still little more than devices that lay down layers of plastic to make three dimensional objects, but they represent a technology that will replace the transport and storage of physical objects allowing them to be composed on site via information transmitted across the network.

Mass customization

Before the industrial revolution, products were produced for individuals. Rifles were made for individual clients as were saddles, clothing and homes. With the industrial revolution we entered an era of mass production. As the saying goes, you could get the Model T Ford in any color so long as it was black. I still recollect the transition from the black Bell telephone at home, which was like everyone else's, to the Princess phones in multiple colors and configurations. It was still mass production, but with new economies of scale allowing different plastic covers.

While the world of products moved quickly to mass production, the service industry was slower. Restaurants were mostly single proprietor operations where customers could be treated individually. Gas stations were full service operations and customer loyalty was valued. Supermarkets were local operations with individual owners who could fulfill your order based on knowledge of your use. Butchers, fish mongers, and book sellers knew their customers and what they liked. Imagine walking into a superstore today and having the butcher tell you they got some great lamb in this week – just the kind you like. There are fewer and fewer special prices, special cuts, just-for-you opportunities in our world.

The automobile industry provides an example of the forces around mass production and mass customization. From the mass-production model of the Model T Ford, the US automotive industry moved in the direction of creating a car for an individual. In the 1950's the demand for cars, and other consumer goods led the US auto industry to try to make cars that were individualized to the needs of each customer. Big cars with thousands of different variations were possible. In Japan, during the same period smaller cars were being made, with fewer independent options. When the first oil crisis occurred in 1973, the cheaper more cost efficient (in terms of manufacture) Japanese and European cars began to get a foothold in the US. Today, following the foreign manufacturers, US automobile manufacturing is returning to limited variability across their models.

In other industries we are beginning to see mass customization emerge with success. There would seem to be little doubt that this mass customization will spread across other industries – including the automotive. The best examples of mass customization include Dell and Amazon. Dell is well known for its ability to construct a machine that meets your particular needs. They begin with base models and change the interchangeable and standardized parts to provide a machine that meets your needs. An important part of how they accomplish this is by providing a customer centric model that allows the customer to indicate what they want to do. The customer then trusts their friendly vendor to provide the options that will most help them. Amazon does a variation of collaborative filtering to cross selling and up selling – i.e. customers who looked at X also looked at A and B and customers who purchased X also purchased Y. These are very different examples of mass customization, but they are reminiscent of the return to a more personal customer vendor relationship – which is surely a trend for the future.

The most significant opportunities for mass customization occur in information intensive products. We have already seen some of this. Various websites are customized to present a personal appearance. Communications to us are personalized – e.g. a utility bill shows usage and temperature for the past year. Books are the quintessential information product. While mass production is still the primary mode, the benefits of mass customization are strong and will impact segments of the industry. In 1986, as part of a research project, we produced custom books to accompany a national television series. Using mainframe computers and a hour or so of CPU time, we were able to produce an integral book to the specifications of a faculty member. Millions of variations were possible, and when the book was done, it had a table of contents, lists of figures and tables, footnotes and references and an index. But each one was different. What cost about $1000 a copy for 500 copies in 1986 could be produced for $30-40 today. We are already seeing a little of this in individually prescribed instruction systems in the elementary grades. The question is why every textbook can't be personalized for every student. Similarly, organizations welcome new employees or customers with a wealth of paper. Whether it is a new employee at a business, or a student at a university, there is no reason why those materials cannot be customized. A student doesn't need the floor plan of every building, just the ones they will use. An employee doesn't need to know the faces, names and responsibilities of every employee, just those with whom they will work.

Pull Versus Push

It is not hard to understand what is meant by push. It refers to the design and production of some product that is then pushed into the market. The mechanism used to push the product is marketing. So, the basic idea is that the automotive industry spends 5-8 years in the design and production of a car that they then work hard to convince people they want to buy. To avoid picking too much on the automotive industry, higher education designs educational programs that it markets to students. Similarly, magazines collect a series of stories that are then marketed to consumers. This is not to suggest that what is being sold is ignorant of consumer interests. Auto dealers want to build what they believe consumers want, just as universities want to offer programs that are designed to meet the interests of consumers. The real difference is that in a pull market place, the goal is to be immediately responsive to informed consumer needs to minimize the need for push, or selling.

Again, we turn to Dell as a simple example of a manufacturer who is less concerned with selling a particular model and more concerned with being prepared to respond to what the consumer wants to pull from them. The notion of structuring an organization or a business that is pull oriented rather than push oriented is sometimes more difficult to see. Consider as a first example how news looks different if it is focused on responding to user pull. A personal pet peeve is the habit of local news stations to tease the viewer. "*Later in the show we will tell you …*." Then they push the news and advertisements we have to listen through to get to what we want to hear. It is a pet peeve primarily because it takes longer to tell me they are going to tell me than it would to tell me. A pull form of news is represented in the many applications that exist to provide very focused news feeds – e.g.

the weather, or the current quote on a stock, or traffic information. In general the use of RSS[56] feeds is an example of a narrow channel that can be watched by an agent to provide access to particular kinds of information.

In the last analysis, the growth of a pull orientation is based on the belief that it is possible for organizations to process information in real time that reveals customer preferences early enough to manage adjustments to the manufacturing process and supply chain such that changing customer preferences can be met in real time. For an institution of higher education this would mean that no course would ever be closed and faculty would be hired and curriculum changed to meet changing student needs. Agile organizations are able to respond to changing market needs more quickly than push oriented organizations.

Use Versus Ownership

This is a complex conceptual trend that impacts organizations and individuals. In the physical world, the growth in automobile leasing versus ownership and the recent decline in home ownership numbers are examples. There are economic benefits for certain classes of individuals in both these situations. Home ownership as a solid financial investment may still be true for many, for others there are better returns in placing money elsewhere. For organizations, paying for services rather than owning expertise has been a trend for many years, but it has begun to accelerate. One of the first services that was outsourced was payroll where it was cheaper for many organizations to buy the expertise as needed from another organization. Progress follows different patterns for organizations of different types and sizes and individuals of different ages and income profiles. Ownership is slowly giving way to pay for use. Several forces contribute to the spread of this phenomenon. The biggest single force is the availability of high speed networks.

For an individual, many new services have become available only on a subscription basis. These include television services, internet connection services, cellular phone and data services. Other pay for use services replace ownership options. These include music, video, and now e-

[56] RSS in common use today stands for "Really Simple Syndication", or a mechanism that can be used to produce narrowcast feeds that users can subscribe to. RSS has an interesting origin in that the original specification, developed by Netscape, was intended as a mechanism that would contribute to the development of the Semantic Web. In this form, RSS stood for RDF Site Summary, and RDF stood for Resource Description Framework. In short, RSS was initially intended as a way for web sites to collect descriptive information, metadata, on all of the resources on the website. More simply still, the RDF site summary might provide information about the resources on a website such as what they were about, when they were created, who authored them, etc. that could easily be collected and indexed by semantic web services.

books. Given a network connection that has been purchased for other purposes and is considered a sunken cost, the cost of Netflix videos delivered to the home is $9/month. The cost of a recently released DVD is between $10 and $20. Ignoring the cost of acquiring the DVD, the cost of a subscription to Netflix is lower. Even considering the multiple replay possibilities of ownership, network delivery allows unlimited replays.

For organizations, there are many issues that impact the decisions to pay for services rather than own expertise. One is the rate of change in products. Thus services that require conformance with dynamic outside forces such as tax laws or network security threats suggest purchasing the services from an organization dedicated to maintaining expertise and servicing large numbers of organizations at a cost that is less than the cost of maintaining that expertise in-house.

Network Directories

Increasingly, we make use of network storage for our information. Google has our mail, Facebook has our friends, LinkedIn has our professional contacts, and Delicious has our bookmarks. This trend is likely to continue and better privacy and security models will emerge. Facebook is closest to a directory of humans, but the nature of the social system is that it is not optimized for look up.

Consider the development of personal certificates for every person that identify them in much the same way certificates are used to assure users that the website they have contacted is indeed the website they intended to go to. Without certificates, or some other authentication mechanism we would never be able to be sure we are providing data to our bank rather than some impostor on the network. Personal certificates would assure that communications from me are actually from me. Certificates could also be endorsed by other individuals or organizations. My certificate could identify me as an employee of the University of Pittsburgh, a researcher who has met Veteran’s Administration guidelines for human subjects research, a member of the Association for Computing Machinery, a friend of Joe Smith, etc. My certificate might also say that I am willing to accept communications from the Institute of Electrical and Electronic Engineers, or from the National Science Foundation. All of this information might be encapsulated in a network directory where other people could find it and submit queries.

The technology of digital certificates, an infrastructure of trusted authorities to issue them, and directories to store them has all been in place for several years now. What has been lacking is a compelling reason for individuals to use them and a good system for managing the complexity of endorsements and conditions. It is unclear what activity will move users to adopt certificates or for organizations to provide really good software that makes certificates manageable and rich. What is pretty clear is that some such online, secure, rich, and manageable store will emerge and become the key to insuring both privacy and accountability on the network.

Smart Objects

There is no doubt that objects are getting “smarter.” When a car can park itself, or determine a hazardous road condition, the device is demonstrating some level of “smartness.” I am tempted to use the term intelligence, and that would be fair in some senses, but the reality is often much less awe inspiring when the covers are taken off. A simple example would be anti lock brakes. Growing up, we were taught to pump our breaks when trying to stop in wet or icy conditions. Why? Because when you press the brakes in a slippery condition, the wheels stop turning and the car skids and cannot be controlled. After many years of research, reliable systems were built that sensed the wheel locking and electronically released and pressed the brake hundreds of times a second to give the driver a maximum level of control and braking. The system needs to understand that it does not get activated during normal slow braking and takes over only during a panic stop when the wheels lock. The combination of simple sensors, microprocessors, and actuators is making a lot of devices more intelligent. Cars are perhaps the best examples of standalone devices that exhibit some level of “smart” behavior from automatic headlights and windshield wipers to anti-lock brakes, airbags, and electronic stability control.

Smart devices are beginning to appear in homes as well. TVs, VCRs, game consoles and other electronic devices are smarter and more self analytic. Hooking up a TV or a VCR to a wireless network takes the input of a few numbers. Most of the process is based on simple assumptions about how a home network works. In several rooms in my house I have installed light switches that have motion sensors in them to turn the lights on when I enter the room and turn it off when several minutes have passed without any motion detected. Outdoor lights go on when night falls and turn off when day breaks. Microwaves sense the steam put out by food as well other chemicals to determine when food is done. With time, they will only get more competent. Refrigerators look like they may be the next device to get smarter. They are already pretty good at controlling temperature and monitoring the state of water filters. When products are labeled with RFID tags (see below), it becomes easy for the refrigerator to help us with the shopping list. All these devices, and many others in the home, will have sensors that detect imminent product failures. They will then make use of wireless networks to send mail to their owner or the vendor. Interestingly, the process of preparing for smart devices is more than two decades old. In 1990, Sun started a project, called the "Green" project to develop a programming language for the new generation of smart devices that were expected to start emerging. As it ends up, the development of these devices was slower than had been anticipated and the new language called Oak was repurposed to take advantage of the growing popularity of the Web – where it became known as Java.

Devices became smarter when it was possible to add cheap sensors, and sometimes actuators, and were able to communicate via small dedicated screens – such as thermostats or via messages transmitted over a wireless network. Objects are also getting smarter via connections to databases. While the history of bar codes dates back to the 1950, it wasn't until 1974 that a system was put together to print a particular style of barcode – the Uniform Grocery Product Code, that could be scanned at a checkout counter to identify a product and look up the cost in a database. The use of the code made pricing and checkout immensely more efficient. Today, barcodes can be printed that include the cost of variable cost products such as meats and fish in supermarkets.

The next generation of object identification will overcome some of the limitations of barcoding which requires the alignment of a scanner and the code. This identification method will likely be some variation of Radio Frequency IDentification (RFID) technology. Active RFID technology was used as early as World War II to identify airplanes as friend or foe. The technology simply broadcasts a signal that can be read. While there are many applications for active RFID, it was not until 1973 that a passive RFID device was demonstrated. Basically, the tag converts incoming radio waves to current which is processed to generate a signal which is transmitted and read, or more simply, active RFID needs a power supply, passive RFID harvests energy from the environment. Over the last forty years, researchers have worked on the development of the technology to include new features and most importantly to reduce the cost to a level near the cost of imprinting a barcode on packaging. There are several advantages of RFID that make this development worthwhile. First, eliminating the need to align the scanner to the barcode means that a grocery checkout could be done in seconds rather than minutes and the need for a clerk would be further reduced. Another benefit related to the ability of RFID devices to be coupled with sensors such as thermometers and accelerometers which would be able to report when a package of chicken got too warm or a package, such as a computer, was dropped.

Perhaps the biggest change we might imagine is the shift from passive to active documents. For centuries, humans have organized and stored documents for later retrieval. It is not too hard to imagine a future where documents would be programmed to seek out the humans who need them! How would that occur? The last section addressed network directories. This would make it possible to find people who had expressed an interest in a given subject, or who were connected with many people who had an interest in the topic. Documents could have active components that sent them roaming the network, much the way spiders do today, placing themselves in inboxes for consideration by the human or agents scanning incoming documents for their owner.

Agents

Agents have been the subject of much research over the last couple of decades. They have been variously defined and take numerous forms. One of the more concise definitions is provided by Patti Maes. *"Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed."*[57] Wooldridge and Jennings provide a longer definition that begins to define some of the important features:

> *... a hardware or (more usually) software-based computer system that enjoys the following properties:*
>
> - *autonomy: agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state;*
> - *social ability: agents interact with other agents (and possibly humans) via some kind of agent-communication language;*
> - *reactivity: agents perceive their environment, (which may be the physical world, a user via a graphical user interface, a collection of other agents, the INTERNET, or perhaps all of these combined), and respond in a timely fashion to changes that occur in it;*
> - *pro-activeness: agents do not simply act in response to their environment, they are able to exhibit goal-directed behavour by taking the initiative.*[58]

A well known class of agents includes the spiders or robots that search the web for documents and were developed to harvest pages for web

[57] Maes, Pattie (1995), "Artificial Life Meets Entertainment: Life like Autonomous Agents," Communications of the ACM, 38, 11, 108-114, page 108

[58] Wooldridge, Michael and Nicholas R. Jennings (1995), "Agent Theories, Architectures, and Languages: a Survey," in Wooldridge and Jennings Eds., Intelligent Agents, Berlin: Springer-Verlag, 1-22, page 2.

search engines. Another class of agents, called bots, includes those developed by various groups to inflict damage on the internet and its users. A set of bots installed on unsuspecting user machines and controlled by their creator is called a botnet. While agents of these two classes will continue to exist, for good and for bad, we are interested in personal agents that will act on our behalf. Still more specifically, we are primarily interested here with task specific agents that operate on behalf of an individual.

Agents are already a part of our computational world. For example, consider the spelling program that is correcting, or suggesting errors in the words I am typing. Ten years ago, a separate spell checker would be run over the text to identify possible spelling corrections. Now, the spelling agent in Microsoft Word operates at both an assistant level and an executive level while I type. As an assistant, it patiently underscores words it thinks are wrong and waits for me to right click on the word. I can then accept its suggestions, ignore them, or add the word to a personal dictionary. But the spell checker also has auto correction options that can be turned off or left on. When the option is on, and one of the spelling situations is encountered, the agent acts in executive mode and makes the correction for me. One might imagine that assistance agents would ultimately develop enough experience to act in executive mode. Spellchecking is rather simple. Many additional specialized agents might also help me. They could include reference checking, searching, appointment scheduling, travel reservations, etc.

Beyond the level of autonomy an agent will have in executing a task, we might wish to control their communication. An agent might be restricted to communication with me – as my spelling agent is. Other agents, especially those who oversee communications, might be allowed to communicate with other agents or other people. Finally, agents might be allowed to coordinate objects, processes, or people, or to control them. When training a spam agent, I initially watch closely to build exception lists and rules. With time, as I find it is making fewer mistakes, I accept the fact that it is acting correctly in 99% of the cases and let it work on its own. In the next chapter, we will talk a little bit more about an example of collaborating agents and surrogates.

As we begin to think about a world of agents, there are a number of things that need to be considered. With a multi-agent system, several additional levels of complexity are added. When does an agent come into existence? When will an agent cease to exist? How do we find agents? How do agents exchange messages? What restrictions are put on agent to agent communication?

Assume agents are loosely coupled software modules that have persistent memory in the form of data saved in files, intelligence in the form of embedded logic, and communications capability in terms of a set of HCI interfaces and socket interfaces. There will need to be an agent directory where agents or humans can check on the availability or existence of a given agent. While the precise protocols will take years to work out, each human and system might have an alpha agent. It will be the job of the alpha agent to keep track of functional agents. The alpha agents will be registered in the directory service and will direct peer agents to the appropriate functional agent or take other actions as required to maintain the flow of communications.

The Semantic Web

Related to agents, particularly in the context of the web is the development of the "Semantic Web." This vision of the web would includes facilities that allow resources, both static and dynamic, to be manipulated algorithmically as well as by humans. The initial design of the Web focused on human navigation of the space. As the space grew in size, and technology developed, it became desirable to develop agents or computational robots (or bots) that could find information for us. A bot is, in concept, very simple. The program begins on a given page which is read into memory. The program does two things. First, it scans through the page looking for links to other pages. Those that it finds, it puts in a normal form and adds to a list of pages to look at – if they are not already on the list. Second, it scans through the page looking to check to see if the page might be of interest. If it is of interest, it makes note of the page and continues searching. This is how the search engines gather and index pages. They begin at the main page of a given site and follow all the links within that site, bringing them back to the search engine for indexing. The vision of the semantic web imagines more sophisticated bots that might find pages with information of interest to their owner and return it for analysis.

As it ends up, algorithmically processing a webpage is not a simple task. It is hard to know that a particular page shows the rates for a hotel in New York City, or the courses offered in US history at a college. So finding colleges near me that offer a course on some topic, or the cheapest rate for a room at a three star hotel in a given city is not a trivial task. Berners-Lee and a number of others envisioned the semantic web as an overlay to the web that would make such searches easier. Doing this involves a couple steps. One is the development of descriptions of web pages that provide more semantic information. Again, simplistically we might imagine a description that says web page X was "authored by Michael Spring", and was "created on May 13, 2003", and is "about the semantic web." The idea is that the critical data would be abstracted from, or added to, a description in a known form that could be accessed by a bot. While at first glance this appears reasonable, there are billions of pages that need to be described and we need to agree about what information should be included. If we were searching for academic papers knowing the subject, author, title and institution might have some merit. On the other hand, if we were searching for cars that had a certain amount of rear leg room or had a certain level of fuel economy, we would need other data.

So, the development of resource descriptions depends on detailed specification of the descriptions. These, maybe inappropriately, are often developed in accord with an ontology. (In philosophy, ontology is the study of what exists and how those entities are related. Information science has appropriated the term and allow multiple ontologies to be defined and used to describe or classify resources.) The basic idea was that as necessary, a group would develop an ontology that would be useful in describing their resources. Given the fact that the web is not centrally controlled, there would be no coordination of the ontologies except to the extent that users felt the need to do so. That would mean that color descriptions of shoes in various languages, or shoe sizes for different regions would have to be mapped, and this gave rise to a significant amount of work on ontology mapping and matching.[59] Further, a lot of work was done on trying to develop automatic and semi-automatic means of developing the metadata.[60] While these activities, and newer efforts to define semantics using "wisdom of the crowd" approaches, with time, information will be overlaid on the Web to increase the level of semantics.

The Social "Periphery"

When we participate in a meeting, go to church, or attend a party, we extend our social antenna to gather social information. We look at body language, tone of voice, eye contact, and other cues to get a sense of the full message being communicated. Looking out at students during a classroom lecture, it is not hard to gauge how much information is being communicated. This information can get lost in digital communication. In an online course, it is difficult to know whether the students are "getting it." There are ways to get such feedback. The feedback can be explicit or inferred. There are both physiological and behavioral mechanisms that can

[59] Mao, Ming, Peng, Yefei, & Spring, Michael. (2010) An Adaptive Ontology Mapping Approach with Neural Network based Constraint Satisfaction. Journal of Web Semantics. 8(1) pp 14-25.

[60] Syn, Sue & Spring, Michael, (2008) Can a system make novice users experts? Analysis of metadata created by novices and experts with varying levels of assistance. International Journal of Metadata, Semantics and Ontologies. 3(2) pp. 122-131.

be used to try to assess this social periphery.[61] Research at IBM and the MIT Media Lab has explored how eye tracking, hand pressure, galvanic skin response, and body temperature might be gathered and interpreted without being intrusive.

As a simple example of explicit feedback, I participated in a teleconference in which the users could indicate what they thought about the pace of the talk – too fast, too slow, about right. As people signed on to the conference, a panel of little squares filled with green dots – arranged as the chairs in the auditorium might be arranged. Just having a sense of how the room filled up as the time approached had an interesting impact on me. As the lecture proceeded, a user could indicate that things were going too fast or too slow. If they hit too slow, the green turned to yellow. If they hit too fast, the green turned to red. Not much feedback, but it was enough. The beauty was the simplicity of the display. It was easy to see people come and go and to get a sense of what they opined about the lecture. There were other features that made the system usable and intuitive, but the basic idea of adding some of the social periphery back in is clear.

In the 1990's we built a system for collaborative authoring that allowed people to work together in groups to create, review and edit documents. The goal was to allow people to focus their energy and to do everything we could to enhance all of the processes such that the process was easier and more efficient. The system was a distributed system and the server kept track of who was working when and what they were doing. (For example, because we tracked all this data, it was possible to tell a team leader what might be the best time for a synchronous chat – which was of course a built in tool. As the system matured, we experimented with an awareness agent to provide information about the members of a group based on activity in a given project and on a given document.[62] We made some inferences from user behavior about how active the user was, how available they were, and how committed they were. Activity was measured by a function that combined edits and comments with time reading the document normalized by document size. Availability was a function of time connected to the system and time expected on a task plus a self assessment of how busy the user was. Commitment was based on a self report plus type of

[61] Barreto, Armando, Zhai, Jing, and Adjouadi, Malek, (2007) Non-intrusive Physiological Monitoring for Automated Stress Detection in Human-Computer Interaction. In Human–Computer Interaction: Lecture Notes in Computer Science Springer Berlin / Heidelberg, pp 29-38.
Picard, Rosalind, (2000) Toward computers that recognize and respond to user emotion. IBM SYSTEMS JOURNAL 39(3-4) ISSN: 0018-8670 pp 705 – 719.
[62] Spring, Michael and Vathanophas, Vichita (2003) Peripheral Social Awareness Information in Collaborative Work. Journal of the American Society of Information Science and Technology. 54(11), pp 1006-1013.

comments made and the overall number of comments made. Each user was shown on a photo bar where the clarity of their image was based on activity, commitment was shown by a tint added to their photo – red for not committed and green for committed. A slide bar showed their availability. Team leaders could see the people tasked on each document and get an immediate picture of who was active, available and committed. This experiment was one early effort to add the social periphery into a digital system.

Over the next hundred years we will develop, learn to accept, and learn to interpret measures of the social periphery. They will not always be accurate, and people will have different levels of skill in making use of them – just as they do in the real world. In some sense, people are already developing this skill, without focused tools in socially intense environments such as Facebook.

Security Assured Information Systems

The internet was designed as a part of a large research effort. Different people define different research goals, but it was clear that the community was close knit, creative and contributory. If malicious behavior existed, it was generally unintended or at worst a kind of practical joke. Most importantly, the network was designed to support the exchange of ideas and research resources. Money was not involved. Today, the internet carries a huge and growing amount of personal and financial information. Given the nature of digital information, particularly the ability to easily manipulate it, it is not surprising that criminal elements have gravitated to this environment.

There are a couple stories about the earliest viruses and worm. My personal favorite is about a worm at Xerox PARC.[63] In 1978, John Shock was working on a program to analyze traffic pattern on the Ethernet inside the research center that connected the 200 Alto personal computers. He was doing some of the basic research on what happened with traffic collisions on the network. To test this, he needed to put a program on each of the machines that would broadcast messages during downtime at night. To avoid having to go to each machine, he figured out how to load them across the network. In a second generation test, he gave each machine the ability to load the program on other machines, autonomously. In theory, when users were working, the program would hibernate. When they left, the

[63] Hiltzik. Michael (2000) Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age. Harper Business, pp 289-299

programs would become active, including the ability to find other machines that were idle and didn't have the program. At some point, the program became corrupted and crashed its host. In the process, it replicated its corrupted self to other idle machines, eventually causing them to crash. When the research staff arrived the next morning, their owners, knowing crashes occurred as a normal part of operations on these experimental machines, rebooted them and started to work. Every time they walked away from their machine, leaving it idle, the corrupted process on other machines would reinfect and crash them. In the last analysis all of the machines had to be turned off to prevent the process from propagating. Imagine a situation like that in today's environment.

One way to think about making the digital world more secure is to draw some parallels to the physical world. Why aren't houses burglarized, cars stolen, people robbed frequently. There are many reasons, but four come to mind. First, we install prevention mechanisms – barriers to access. Houses and cars have doors that can be locked. Cities have police that patrol the environment, people can carry weapons from pepper spray to guns. Equally important, these mechanisms can be installed at varying levels of strength given the hostility of the environment and the value of the assets being protected. Second, we have alert mechanisms to mitigate the process – house and car alarms, etc. Third, cultural norms exist related to physical property. While they are not honored by all and while situations can reduce adherence to the norms, as a culture, we generally respect that ownership is not to be violated. Again, accepting all of the caveats, the vast majority of people would never consider stealing some item from a house they were visiting, or using someone's car without their permission. To a great extent, this cultural restriction does not exist in the world of digital information. Many individuals who would never consider taking a audio CD from some valid owner have no problem making a digital copy of songs. Fourth, the risk to benefit ratio is much higher in the digital world. To rob one person in a park entails a certain level of risk and has the likelihood of providing a reasonable benefit. In the digital world, the risks that a perpetrator faces are much lower and the potential benefits are much greater. While laws and penalties are catching up, they still lag and the benefit of gaining access to the money of thousands or millions of individuals is irresistible.

So, how do we make the environment more secure and safeguard peoples assets? Here we address four components:

5. Better infrastructure
6. Better programs
7. Better users
8. Better laws and enforcement

We discuss these trends in greater detail below.

Better infrastructure

The infrastructure consists of both the network and the computing platform. The network must be secured at a number of levels. Network reliability, survivability, and scalability insure that messages will be able to get from a source to a destination. In other words, access will not be affected by catastrophe, network load will not impact performance, and malicious operators will not be able to subvert the infrastructure. Efforts in this direction have been underway for more than a decade. The details of how it works are less important than the concepts, but basically a protocol known as Transport Layer Security(TLS), which replaced the Secure Sockets Layer(SSL) protocol developed by Netscape in 1995, provides a kind of tunnel which provides for the encryption of the messages between two endpoints on a network. This addresses the issue of transactions and involves the use of a digital certificates and signatures. Most non-technical users can see this process in action when they log in to a commercial website like a bank or an online business. Before any information is passed back and forth, you will see a change in the address of the website. The first part of the URL, which is normally "http://" will change to "https://" which is an indication that the communication with the web server is occurring with transport layer security. A connection to the main University of Pittsburgh website is not secured, but access to the internal portal is.

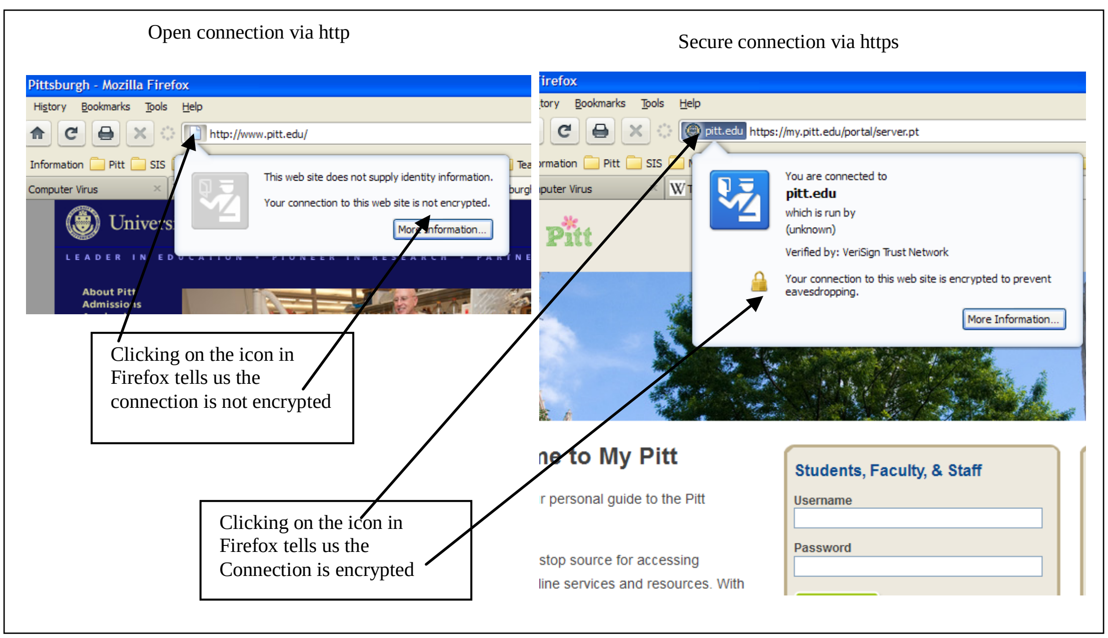


Figure 27: Normal unencrypted and TLS connections

Providing a secure connection between two machines is a big part of the process of securing the network transactions, but it is far from sufficient. In 1995, in response to concerns about security, it was suggested that rather than building a tunnel, it might make sense to encrypt every packet. This would be security at the level of the basic packet protocol – which is the Internet Protocol (IP). At the same time IPsec was emerging, the community was engaged in discussions about expanding the number of devices that could be addressed on the internet. The shift from the 32 bit addressing of IPv4 to the 128 bit addressing of IPv6 was complicated by several factors – not the least of which was that all of the devices on the internet had to be able deal with the new addressing scheme. While IPv6 with security will go a long way to securing the internet, it has not become adopted as quickly as might be hoped. Another major network security issue relates to the Domain Name System (DNS). In place since 1984, the DNS provides a kind of distributed phone book that associates names with IP addresses. Thus, rather than having to remember that the Pitt website can be accessed on about 6 machines with addresses such as 136.142.188.76 or 136.142.188.73, we simple type "www.pitt.edu" and the domain name service provides the address the machines need. DNS, like the other internet protocols was written at a time that malicious activity was not a concern. It is not hard to imagine the havoc that might be created if the system was corrupted and your request to be connected to your bank was hijacked and sent to an impostor machine controlled by criminals. Beginning in 2005, serious efforts have been underway to add security extensions to DNS to close up these vulnerabilities. Like so many other security efforts related to the internet, the task is made more difficult because it is a big system with millions of locally controlled parts that has to be changed. Otherwise the whole system breaks down.

The second part of securing the infrastructure involves securing the business endpoints that are connected to the internet. Again, the issues are complex and involve hundreds of thousands of professionals. Basically, businesses need to be placed behind barriers that prevent unauthorized access. These barriers are commonly called firewalls. The basic principle is that machines that have sensitive data are positioned such that they can only be accessed in very carefully constrained ways. Each machine connected to the internet can be accessed using a number of different programs, each mated to a server process. For example, I can connect to one of my web servers by using a program called ssh – Secure Shell. If someone steals my username and password, they can connect to my web server and take any data I have stored on it. To avoid this kind of attack, I should be prevented from connecting to my web server over the network. (To change something on the server, I would have to go to the locked room where it is kept and log in on the machine's console.) This would prevent someone who stole my username and password from accessing it remotely. Indeed, in an environment of tight security, each asset would be allocated to a separate physical machine. Thus, if someone stole my user name and password and accessed my web server, they would still not be able to access my database server which would be yet more isolated. Servers of this kind are placed on host machines that are called bastion hosts.

Great progress has been made over the last decade in building and securing bastion hosts and placing them behind firewall. Methods have further been developed to monitor these machines and networks to detect and react to intrusions. These intrusion detection systems work in a variety of different ways, but basically they are structured to detect any anomalous behavior. All of these efforts to create environments that are resistant to unauthorized access fail however when the access is authorized but malicious. With effort, machines can be kept safe from criminals on the outside, but they are still vulnerable to internal attack. This can be as "innocent" as leaving a terminal with access to private data unattended. It can be as "criminal" as an individual downloading data during an approved connection and then physically moving it somewhere where it can be sold to individuals will to pay for it. An eight gigabyte flash drive, costing ten dollars, can store the records of millions of individuals that might be worth hundreds of thousands of dollars.

Better programs

Even when the infrastructure is appropriately secured, there are security vulnerabilities that have to be addressed through better software controls. In the past, and still to this day, few academic preparations prepare students to build robust and resilient systems. Indeed a major boost to improving the quality of software came from Ronald Reagan's "Star Wars Defense." Reagan announced the Strategic Defense Initiative in 1983. One of the obstacles to being able to build such a system was the need for software placed in space that might not be used for years and then would have to function flawlessly the first time. There could be no software glitches. In 1984, the federal government created the Software Engineering Institute (SEI) as a federal research and development center to advance software engineering and the development and operation with predictable and improved cost and quality. The SEI became home to the Computer Emergency Response Team (CERT). The role of CERT has grown and changed over the years. It used to be active in the analysis of and response to attacks. Today, it is more of a research center and works to assist the many organizations now dedicated in organizations and across nations to insuring the security of the internet. It remains, as a part of the SEI, a location for research of security vulnerabilities.

So, what are the kinds of programming errors that cause problems? There are about two dozen major categories of errors related to various programming processes such as string manipulations and dynamic memory management. We will describe two with apologies in advance to those who know the details for the facts that we will leave out. The first is the "call stack buffer overflow." While the details are far more complex, it should not be hard to understand that a program is run by executing one instructions after another that are place in memory. So a program might conceptually look something like this in memory:

Address	001	002	003	004	005	…	110	111	112
Contents	Run	Hop	Skip	Add x+y	Hop		Stop		

The computer starts at address 001 and "runs" and then moves to address 002 and "hops", etc. So far, so good, but we need to store data in a program, not just instructions. So let's imagine a slight change to memory as shown below.

Address	001	002	003	004	005	006	007	…	112
Contents					Hop	Get Name	Stop		

Now we have both a data storage space – from 001 to 004 followed by instructions. The computer starts at the instruction at memory location 5 and continues from there. When it gets to the instruction "Get Name" it fetches the name provided and writes it starting at position 001. Most programmers would anticipate the size of the name to be entered, and make sure to leave enough space. So, if Pat entered his name, memory would look like this after the name had been gotten.

Address	001	002	003	004	005	006	007	…	112
Contents	P	a	t		Hop	Get Name	Stop		

But let's imagine that Jonathan decided to enter his name. Because his name is larger the space allocated, unless checked by the programmer, the result can be a buffer (space) overflow. It would look something like the following:

Address	001	002	003	004	005	006	007	…	112
Contents	J	o	n	a	t	h	a	n	

Note that the instructions in memory have been overwritten by Jonathan. Now, this presents a problem in that the processor is going to get its next instruction from memory location 007, which should be "Stop." Instead, it is going to find the letter "a" and then the letter "n." If the person entering their name was very smart, what they would do is make up a name where the "letters" were the operations codes that actually stand for instructions in memory. (Keep in mind that a set of zeros and ones in memory is interpreted differently based on what the computer expects is stored at that location. Writing a string that could be interpreted as a program is, in the last analysis, not all that hard.) A "Call Stack Frame Buffer Overflow" is a special condition that allows the hacker to take control of a program for his/her own purposes. Figure 28 included a small piece of an html page with a script.

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en"><head>
<title>Homepage for Michael B. Spring</title>
</head>
<body>
<h1>Michael B. Spring</h1>
<p>You have reached the homepage for Michael Spring, a faculty member in the
Graduate Program in Information Science at the University of Pittsburgh.</p>
<p>These pages present information about teaching, research,
 publications, as well as information about my background and personal interests.</p>
<center><script>
      var theDate = "";
      theDate = document.lastModified;
      document.write("This document was last modified  <br/>");
      document.write(theDate);
               alert('boo');
</script></center>
</body>
</html>
```

Figure 28: HTML Document with Script

A simpler programming error relates to a technique called cross-site scripting. It begins with requiring an understanding of how browsers work. When we load this page into a browser, what you would expect would happen with the paragraph will happen. The interesting part is what happens with the script at the bottom of text. When this page is loaded, **Error! Reference source not found.** shows what appears in the browser.

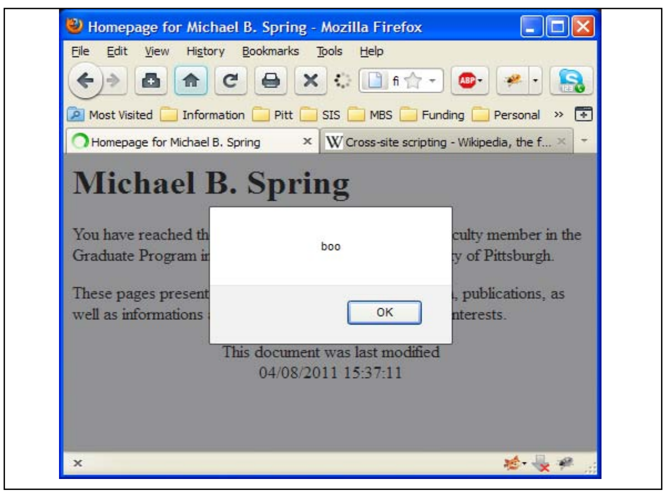


Figure 29: Alert Box

You will note two things. First, the date of last modification of the file has been filled in by my little program or script. Second, a window has appeared over the text saying “boo” and demanding the user acknowledge it. Pretty intrusive. But now we get to the real threat. Imagine that some site that shares data about users asks me to input my name. I input "Michael<script>alert('boo');</script>." This text string, if unchecked, will be returned to the server and stored. When someone on another machine asks about me, the server will send them my name. My name will appear as it should, but the script is intended to be invisible and it is, but the user will see an alert box with the word "boo" in it. Thus, in this simple example I have used the server to pass malicious code from my machine to some other machine. If data is not checked and cleaned, cross site-scripting can be a powerful tool. There are many more ways that programmers can protect users and protected machines from malicious code, and we are making progress.

Better users

The third leg of improved security rests on the shoulders of hundreds of millions of users like you and me. As it ends up, millions of end user machines have been infected by various forms of "malware." The AntiPhishing Working Group regularly tracks infected machines and while the numbers are improving slightly, in the second quarter of 2010, more than half of the machines checked were infected![64] The kinds of infections vary from code to steal passwords to code to make the machine a soldier in a botnet. There are a variety of tools that can be used to limit attacks, and while nothing is a guarantee, informed and educated users can take simple actions to limit their personal vulnerability as well as the security of the network as a whole. Without these vast numbers of infected machines, many of the exploits being used by criminals would not be possible.

To assist normal users in securing their endpoints, vendors have made decisions to lock down machines and set security levels to high by default. Users can, of course, fail to update the software that checks for malicious programs, or indiscriminately override protections.

Better laws and enforcement

In general laws and enforcement lag behind technology. While seat belt technology is almost as old as the automobile, they were not offered as options in US cars until the 1950's. It was not until 1963 that they were required equipment in cars (front seat). In 1984, laws mandating that seatbelts be worn began to appear, but they still vary by state.

[64] Anti-Phishing Working Group (2010) Phishing Activity Trends Report, 2nd Quarter / 2010. Retrieved May 2011 from the APWG website: http://www.antiphishing.org/reports/apwg_report_q2_2010.pdf, p. 10.

One early effort to update laws relates to copyright. The Copyright Act of 1976 was brought up to date by the Digital Millennium Copyright Act of 1998. Laws related to digital information and security issues have appeared more slowly. There is growing attention being paid to cybercrime. The first major computer security related law was the 1986 Computer Fraud and Abuse Act. It dealt with protected computers – ones owned by a financial institution or the US government. It was under this law Morris was prosecuted for the Morris Worm. Other laws have looked to provide for individual privacy in electronic communications and protections from spam. Recent efforts have also been directed to lurking and the exploitation of children.

Interfaces

The last decade has seen the emergence of more automatic interfaces on the device side and more intuitive interfaces on the human side.

Automatic Interfaces

Homes are moving to some standardized networking interfaces and more and more devices are connecting to it. The Universal Serial Bus (USB) interface replaced RS-232 as the standard interface for peripherals. While software to control the interface still needs to be loaded, frequently a standard driver is already available on your machine, or the device is capable of an automatic bootstrap loading. If software needs to be loaded from external media, the process is generally automated. Given the “smart device” nature of the attached peripherals, the kinds of communication that take place across the interface are increasingly “adult.” That is, they are couched in human terms rather than device terms. “The printer is out of paper” or the “Red toner level is low.” A multitude of specialized peripheral device can now be used by individuals with little knowledge of the technology.

As little as a decade ago, doing research on remote medical service delivery, it took hours to connect cameras and process the incoming data. Today, you plug in a camera to a USB port and it works flawlessly, automatically accommodating low lighting levels, focusing on the objects in the field of view, and following my face if I wish it to. On the internal side, it is linked into Skype to make video calls as well as Word for inclusion in my documents and other software that might make use of images. Most users are also familiar with Bluetooth technology which is the wireless equivalent of USB. Many users associate Bluetooth devices with their cellular phones, but they are also available for PC’s. Teaching in class, it is useful to have a little mouse that is connected to my laptop via Bluetooth. It allows me to walk across the classroom and still control the PowerPoint.

Similarly, if I wish to record an audio annotation for a classroom lecture, I attach a Bluetooth microphone that can serve as the recording source. Interestingly, the Bluetooth microphone replaces an earlier wireless microphone that was designed for remote connection to video cameras that cost 10 times as much.

Beyond peripherals, digital information devices are increasingly connected by wireless LAN technologies. TV and Disc players are now included as devices that operate on these networks. The ease with which the networks can be set up and devices attached continues to get simpler. Each year sees a vast improvement in the setup ease, with the most recent Blu-ray disc player and tablet PC complete all the steps except the input of the encryption key. Similarly, my cell phone requires only a passphrase to be connected to the home wireless LAN. With the connection open, there are apps to control various other devices in my home that are accessible via the network. I can use the phone to control my TV as well as the DVR player. Indeed I can use my phone to connect to and program the DVR from anywhere where I can access the internet. Applications are emerging to make this kind of remote control far more prevalent. Soon, the connection between the digital devices in my home and the digital devices I carry will be so seamless that it will not be any kind of technical challenge to wake up cameras in my house to see what is going on or to have the cameras in my house alert me when they sense motion in the house.

The final step in the home digital interface will be the addition of other wireless networks, both high bandwidth and low bandwidth. The Ultra-WideBan (UMB) standards have been controversial, but various standards have existed since 2006 and some form will emerge as the preferred standard in the near future. Once manufacturers begin to adopt it, we will see a Bluetooth like system that will enable the connection of peripherals requiring high speed data transfer over short distances. We will see a progression similar to that for RS232 → USB → Bluetooth. This progress will see the evolution of interfaces from the composite video, component video, coaxial cable mix → High Definition Multimedia Interface (HDMI)→ UMB. We can expect that same process of simplification, ease of set up, and ease of use. The days of dozens of cables connecting our multimedia devices will be over. More importantly, it is likely that various devices will switch components on an ad hoc basis. Sometimes, the large screen will be driven by the Tuner, sometimes by a DVD player, and sometimes by my computer. The low bandwidth interfaces include ZigBee and Z-Wave. They will be used to control simple electrical devices such as light switches. It has been possible for a decade to include sensors in light switches. Indeed a single pole motion sensor light switch that fits the standard sized light switch space is virtually the same price as a regular light switch. These devices can now be given low bandwidth wireless connections to a control center that will allow me to turn the lights on in my house, or adjust the thermostat as I drive home.

Intuitive Interfaces

Just as device side interfaces will continue to be more automatic and self adjusting, human interfaces to devices continue to get more intuitive. We are not quite to the level of the kinds of interfaces shown in the movie "Minority Report" but we aren't far away either. If we had to pick just one winning technology of the past few years, surely the multi-touch interface would have to be close to the best. Just a few years ago, a touch screen interface required extensive preparation and was limited to a single finger that was not very accurate. A single small screen cost as much as a PC. Following the introduction of multi-touch technology in the Apple I-Phone, I-Touch, and I-Pad, multi touch technology has emerged as the next way to communicate with devices. For dealing with multi-media it has obvious advantages, but there are several places where it is applicable. As a minor extension of this, accelerometers now allow devices to switch orientation as naturally and quickly as one might want.

Voice and handwriting are also making progress as natural and intuitive interfaces to devices. For tablets, handwriting can provide a significant advantage in some applications such as note taking. Voice recognition has a long and tortured history, but beginning with cell phones and other limited use applications, it was possible to achieve a reasonable level of accuracy. The most recent dictation systems have achieved a good level of accuracy with a minimum of training. There are certain disadvantages to these two methods of input. Obviously, speech recognition might not be appreciated for note-taking during a speech and it doesn't work well in a noisy environment. Handwriting tends to be slower than typing depending upon the complexity and length of the message. Voice recognition for dictation requires little skill on the part of the user beyond relatively simple action versus dictation commands. Controlling a device can require significant training and constitutes a load on memory. (The GPS system in my car responds to voice commands and can control the heating and cooling, the radio and DVD, and the map functions.) Unfortunately, with the exception of "I'm hungry" which shows nearby restaurants, the commands are not very intuitive. Because I haven't committed the commands to memory, the potentially impressive system is not used very much.

Voice command technology will likely be combined with eye tracking or some similar technology to make issuing commands, and remembering them easier. For example, there are many things that might be opened or closed, and many situations where next and previous make sense. If I could look at objects and keep a few easy commands in mind, voice activation will make more sense. Other intuitive interfaces will continue to emerge. Successful ones will be copied and poor ones will be abandoned. The automotive industry continues to excel and make improvements in this area. Some cars can now parallel park themselves. Others can sense a face where the eyes are closing while driving and alert the driver to the danger. Still others can keep track of vehicles in blind spots and warn the driver, but only when a complex set of actions indicate the driver is about to switch lanes. All of these intuitive interfaces and others that will be developed will be made smarter by a continued stream of data about the user. We are creatures of habit and we are purposeful creatures. What I will type next in this book is not random but dependent upon all that has preceded. As interfaces make use of this information, they will no doubt continue to do a better job. At some point, fulfilling the dream of Xerox PARC Researchers who first put the key on the development machines in the early 1980's, it will be more likely that we will be able to say DWIM[65] – "Do What I Mean", and have the system correctly execute our wishes. Better yet, like the automatic spell correction we have now, the system may just be able to do it without having to ask us.

Final Thoughts

It is hard to say what is going to happen over the next hundred years. Many individuals will develop viral ideas like the initial web or the current social networks that will only be seen as obvious after the fact. What is undeniable is that the online world is here to stay. More and more of our experiences will take place in digital form or have a digital equivalent. This digital information will make it possible for us to provide more personal and humane services – to pay more attention to customers, be they students, patients, car buyers, or political constituents. It is also clear that objects will get smarter – my car will know where it is and in that world of smart objects, computer agents will be developed to assist us. They may be virtually invisible – like the spelling checker, or more interactive asking us for information to help them do their job. In all of these activities and developments, the most important point is that once we discover how to do something, we don't lose that knowledge. (The voice recognition system in luxury cars last decade will appear in all cars next decade.) While criminals will continue to be attracted because of the lucrative opportunities, we will build better and more secure systems and users of the digital environment

[65] The Alto, which lead to the Xerox STAR, was both a prototype and a development environment which included programming in LISP. The machine included a facility for telling LISP to do what I mean, not what I say. Lyle Ramshaw, The Alto-Dolphin-Dorado Briefing Blurb, or Exploring the Ethernet with Mouse and Keyboard, May 1981, http://www.bitsavers.org/pdf/xerox/parc/Exploring_the_Ethernet_with_Mouse_and_Keyboard_May81.pdf

will learn lock their car and not to walk down dark allies alone. The social periphery is one important area where new developments that are hard to predict are likely.

Chapter IX: Individual Perspective

Introduction

This and the next chapter describe some of the changes that are occurring in different phases of our life. The material might have been organized in any number of ways. The approach I selected looks first at how things will change for us as individuals and then turn to our social experiences. We start with the control of personal information as we look to the future. The chapter then looks at new forms of control and power we will gain including a look at personal agents. The chapter then provides some speculation about how the changes will impact important activities – education, medicine, leisure and work. The chapter concludes with some thoughts about the exposure of our private and public information.

Control of Personal Information

Private Personal Information

Around the beginning of the millennium, Gordon Bell and colleagues at Microsoft began a research project on providing access to all of a person's information, accrued over a lifetime. The project, called Life Bits[66] has resulted in a number of interesting research sub projects. One of those is a projection that the storage needs of a relatively active individual would be about a gigabyte a month, or a terabyte over an eighty year lifetime. When the project began in 2000, the cost of a terabyte of storage was about $10,000. In 2010, the cost of a terabyte of storage had dropped to under $100! If one excludes video, images, and music, the total get much smaller. I have not recorded or saved all my communications, but a

[66] Bell, Gordon & Gemmell, Jim (2007) A Digital Life, Scientific American, March 2007.
Gemmell, Jim, Bell Gordon & Lueder, Roger (2006) MyLifeBits: a personal database for everything, Communications of the ACM, 49(1), pp. 88-95.

reasonably comprehensive archive of notes, papers, books, selected emails is a mere 32 gigabytes. Photos and music add another 50 gigabytes. Movies and Video, all digitized add another 168 gigabytes. The lifetime of my children in video and still photographs, can be passed on in digital form for less than $20!

Transfer of the Super 8mm film, and the analog video, slides, negatives, and prints took more than a half a decade of leisure time conversion. Today, it takes no time. Conversion of floppy disks and other storage media over the years took some significant effort. Now, everything is duplicated across several hard drives to make sure it is up-to-date. In the future digital image resolutions will continue to eat up storage space. Whether, like Bell in the LifeBits project, people will choose to snapshot all web pages they visit gobbling up storage space, it is pretty clear that the dropping cost of storage will make it affordable for every person to make as complete a record of their life as they might desire. And it will be easy to do so.

Beyond the ease of acquisition and storage, we have random access to the information. Long gone is the day when we need to watch an hour of video tape, or several reels of 8MM film to find the scene we want. Once digitized, the various forms can be easily accessed at any point. Further, indexed access tools are growing in power and simplicity of use. Digital camera not only add date and time stamps, but increasingly include geocoding. Microsoft PhotoGallery automates the process of face recognition in photo management. It quickly identifies faces in the image and with just a little effort guesses the correct names to be associated with the facial images of photos it sees for the first time as shown in Figure 30. It is not hard to imagine a little bit further down the road where all sorts of clues will be observed and coded into a tag – such as Christmas trees or Easter Eggs. The LifeBits project is researching other kinds of metadata that might be captured or inferred. If the longitude and latitude are known, place names might be added. Once the face of my children is pointed out once, the system might encode that information in that photo and all the subsequent photos – twins might present a problem, but face recognition technology is very good even today. Emails and other documents can be scanned and keywords extracted to provide better indexing.

Figure 30: Automatic Face Recognition

Network Based Personal Information

We keep increasingly greater amounts of personal information on the network. The amount of personal information stored on the network will likely increase in the coming years with increasing benefits and some risks as well. The kinds of personal information we keep on the network can be separated into three broad groups. While useful for explication, there are complicated overlapping forms. We look at public, private, and semi-private personal information.

Public Personal Information

The basic mechanism for making personal information public is to create a website. Early on, this was a daunting task that required access to servers to write and link a variety of pages. Working in academia, and focused on document processing, it only made sense for me to experiment with the technology as soon as information began to be published. Understanding the protocol, a rudimentary server could be written in a couple of days, but of course, reliable and robust implementations were more difficult. Browsers, especially graphical, were a significant chore. Thanks to NCSA,[67] by 1994 servers and browsers were publicly available and programming was not necessary except for CGI programs which were simple by comparison.

In 1995 as my course on document processing was beginning to include aspects of the web, I developed my first site which provided access to a little information about me. In 1997, we mounted websites for various research projects and in 2001 I removed course resources from my main website and constructed a more interactive website for course materials. In 2007 I added a blog to my website and began to provide periodic postings on various topics. For most individuals, these kinds of resources were not easily accessed, and even though writing web pages was a relatively simple task, it was not something most people had a need for, an interest in doing, or the time to do it. Simpler systems were needed.

Two efforts emerged on opposite coasts in 1995 to provide access for people on a shared basis. The Beverly Hills Internet web hosting service started in 1994 and emerged as GeoCities in 1995. It allowed users to build a web page that could be placed in a city. In 1999, Yahoo purchased Geocities, and for a variety of reasons started the decline of the service. On the east coast, some students and faculty from Williams College started thinking about a similar service. They purchased the domain name

[67] Katz, Eric, Butler, Michelle, & McGrath, Robert (1994) A Scalable HTTP Server: The NCSA Prototype. Computer Networks and ISDN Systems, Volume 27, pp155-164.

Tripod.com in 1994 and opened the public service in 1995. It provided simplified tools for building a resume and a homepage. Today, most organizations that provide internet access provide simple automated tools and space for subscribers to mount and maintain web pages. Three additional services emerged later to provide very different kinds of services for the exposure of personal information.

A web log or blog is basically a set of diary entries In this sense, blogs have been around for as long as the web has existed. Two events gave rise to the modern form of blogging. Efforts to develop semantics for the web created and XML document form called an RDF Site Summary (RSS). Conceptually, the document form is very simple. An RSS document contains a "channel" that has several descriptive elements and includes a number of items. The items have a title, description, link, date, etc. and there may be any number of items. This document schema never caught on as a method for describing a site, but some folks recognized that it could easily be used to contain the elements of a blog. This standardized document format was important in that it made it possible to develop tools to harvest the information that was stored in this format. It made it possible for users to subscribe to a blog and get updates when the blog was added to. The RSS format has given way to ATOM which was designed explicitly as a syndication format.[68] The second development was the creation of a set of websites that made it possible to build these documents very simply. In 1998 and 1999, Open Diary, LiveJournal and Blogger all opened their doors. Besides providing easy editing tools, the services provided for commenting and blog management. Today, blogger has become one of the most frequently visited sites on the web. (In 2011, according to alexa.com, blogger.com was consistently in the top ten sites visited!)

In 2003, LinkedIn, while it has social networking features, also has strong characteristics of more traditional websites. You can share updates with individuals with whom you are connected, and you can manage your network connections. While not all information about participants is provided for non-paying users, paying users generally have full access to information about the individuals and organizations who participate. While it is possible to restrict the visibility of various pieces of information to your connections, by default information is shared more broadly. As the LinkedIn privacy statement puts it:

[68] Nottingham, M. & Sayre, R. (2005) The Atom Syndication Format. Retrieved May 2011 from the IETF website: http://tools.ietf.org/html/rfc4287

> *Being part of LinkedIn means sharing information about yourself with other professionals, communicating with them, as well as working privately on your own. By default, your account is set up to share the information that we have found the vast majority of our Users are interested in sharing. But the amount and type of information you decide to share, and with whom you share it, is up to you.*[69]

The latest addition to the public information publishing phenomenon is YouTube. YouTube allows users to publish personal videos and has quickly grown to be one of the most popular sites on the web. While YouTube does make provisions for restricting access to your video (private videos) or not listing it (not as strong as private if others share the link) the intent is to provide public assess which accounts for the majority of postings.

> *We created YouTube to make it easy for people to share their videos with the world. That said, sometimes you may not want everyone to see your video or your channel, and would rather share your work with only a small group of friends or family.*[70]

Private Information

The most private information that we keep on the network consists of financial and other confidential information – phone numbers, mailing and email addresses, etc. Any time you login, or use a benefit card such as common in supermarkets today, additional private information about you is gathered and stored on the network. Amazon knows what products you have purchased, as does a supermarket or a discount store that requires a membership card. In exchange for allowing the supermarket to record what I buy, I receive special discounts and other benefits. People are selling the right to access this kind of private information at an alarming rate. (I know of no exposure of purchasing information such as that gathered by supermarkets that has been exposed in such a way as to harm an individual.) At the same time, it is not hard to imagine some pattern of purchases that might prove embarrassing to some public person if they were revealed.

Other private information stored on the network includes information that we may or may not consider confidential. All email is

[69] Linkedin (2010) Privacy Policy. Retrieved May 2011 from the Linkedin website: http://www.linkedin.com/static?key=privacy_policy

[70] YouTube (2011) Privacy at YouTube. Retrieved May 2011 from the Youtube website: http://www.youtube.com/t/privacy_at_youtube

passed across the network. Historically, the earliest mail system delivered that email to a local machine when the user connected. The last several generations of email left the user the option of leaving their email on the network. In the initial Post Office Protocol (POP) the default was to have all the email delivered to the user and deleted from the network. Most recently POP and the newer Internet Message Access Protocol (IMAP) make the opposite assumption – that mail would stay on the network. Historically, email accounts were provided by work organizations and Internet Service Providers(ISP). In general, your communications at work were considered the property of the organization with reasonable assurances of privacy. Mail provided by the ISP was generally considered to be private and personal, although there were few laws and policies in the early days. Over the last several years, Google's provision of Gmail has grown tremendously in popularity. While Google appears to be operating in an honorable fashion, users should be clear about what they have the right to collect and use:[71]

> ***User communications*** *– When you send email or other communications to Google, we may retain those communications in order to process your inquiries, respond to your requests and improve our services. When you send and receive SMS messages to or from one of our services that provides SMS functionality, we may collect and maintain information associated with those messages, such as the phone number, the wireless carrier associated with the phone number, the content of the message, and the date and time of the transaction. We may use your email address to communicate with you about our services.*

Beyond the storage of personal communications, users have found it convenient to store all sorts of information on the network. Most of these are directed at easing cross machine access, sharing with friends or both. Calendars and bookmarks are examples of information that is frequently shared across machines. Documents and images are examples of information that we may want to share among individuals. There are systems such as Delicious that allow us to achieve both goals. These systems, designed for sharing as a primary goal, are discussed below.

[71] Google, (2010) Privacy Center. Retrieved May 2011 from the Google website: http://www.google.com/intl/en/privacy/privacy-policy.html.

Semi-Public Information

Social networking sites have emerged as the newest phenomenon on the web. Just as LinkedIn shares the features of social networking and online public information, the sites that would give birth to today's social networks include a variety of different sites. Dating sites such as Match.com which was started in 1993 shared this notion of services which people joined for the purpose of making information widely available publicly. Classmates.com started up in 1995 initially restricted to sharing email addresses and expanded to include profiles in 2000. e-Harmony.com started in 1998. While the dating services are not directed toward communities, they do serve a particular community and include the notion of profiles and personal information.

One of the first social networking sites to emerge was the MakeOutClub in 2000. It was written by a web designer in Boston, and while it still exists, it never emerged as a major force. In 2002, Friendster emerged on the scene and grew rapidly over the first year. Observing Friendster, some employees at an internet marketing company, eUniverse, decided to create a similar social networking site. Called MySpace, it was launched in 2003 and until 2007 was the most successful social networking site. Facebook, which was introduced in 2004, lagged behind MySpace for several years(2004-2008). In 2009, Facebook membership grew fourfold while MySpace stayed level. There are a variety of differences between the two sites, but users most frequently reflect on the fact that Facebook was explicitly designed for college students while MySpace tended to cater to teenagers. The other significant difference between the two sites is the number of applications available on the sites – there are many more on Facebook.

The kind of information shared on social networking sites includes basic personal profile information and photographs and news about what we are doing. There are a variety of controls that allow that information to be selectively exposed to various social groups to which we claim membership. Various applications, games, and posting devices allow us to communicate easily with individuals and groups at different levels. We will address more aspects social networks in the next chapter,

Day to Day Activities

Control of Devices

Today, we are imposed on to use many different devices with very different interfaces. Many years ago, researchers at Sun and MIT anticipated this situation. Sun's response was the Green Project for smart devices that was discussed briefly on page 136. MIT's response was the X window system which allowed processes running on one machine to display information on another machine – and to accept directions across the network. Today, there are various remote desktop applications available for most operating systems. Most people tend to have a hard time interacting with remote machines and processes. (When a vendor remotely accesses a PC to diagnose a problem and fix it, people talk about the process as if it was magic. Despite this, somehow the notion of scheduling your DVR from your cell phone or office computer is easier to understand. Particularly young people view the control of the digital video recorder (DVR) on their home TV system from a PC or cell phone to be a natural activity. In the future, remote control of devices and processes will be common and accepted. Just as I can remotely program my DVR, there is no reason, other than programming and security, that I can't access my computer system at home from my computer system at work and move data or print some information.

The thermostats in my house are already pretty smart – allowing for several weekday and weekend periods. The addition of programming linking the clock and a motion sensor should allow the thermostat to gather data for a couple weeks and then recommend several programs to me based on data it gathers. Do you want energy savings? Do you want above or below average temperatures? Once set, it should accommodate early arrival home, or a vacation as indicated by motion or lack thereof. Further, not that I won't trust them, they should be subject to remote control so I can override any settings. Today, with wireless LAN technology, this might add ten to twenty dollars to the cost. With Zigbee, it is likely that there would be a negligible difference in cost to add remote control. Just imagine all the other devices at home that you might be able to automatically micromanage. The combination lock on the door might be remotely adjusted to allow workers in without giving them your regular access code. You might be able to remotely open the garage door to allow lawn equipment to be repaired or accessed. The lights could be turned on or the door unlocked for company when you are running late. The oven, grill, breadmaker, or crock pot might be remotely started such that they would be ready when you got home. The possibilities are endless. What is not yet known is what people will find useful and what they will find too complicated.

Cars are another easy target for remote control. Just as the cost of a remote microphone (see page 151) drops dramatically when integrated with an existing infrastructure, so the cost of remote starters and other remote control features will become cheaper to add to cars as they become connected to the internet. If my calendar identifies a sporting event at a remote school for the children, the destination address can be downloaded to the GPS for me. But cars bring us to the next topic – ease of use.

Ease of use

One principle of good interface design is that it is as intuitive as possible. While many of us have learned to use scroll bars and mice effectively, they pale in comparison to multitouch technology which requires virtually no learning curve. Given the development of the base technologies, such as multi-touch, interfaces will get easier to use and will converge on common successful patterns. Cursor movement has a role, mouse technology has a role, and multi touch technology has a role. The use of touch screens should make a whole series of small devices more intuitive to use. These will include cameras, thermostats, radios, and a range of other small devices. Even without multitouch screens, interfaces that couple buttons with messages can be very intuitive. Press the beverage button on a microwave and it says 4oz. Press it again and it says 8oz and again and it says 12oz. One more time and it goes back to 4oz. Pretty smart, pretty intuitive.

Another principle of interface design is that good interfaces are invisible. The automotive GPS system I mentioned earlier is a good example at one level – showing restaurants when I say "I am hungry." Unfortunately, few of the other commands are as intuitive. I would have preferred the command "I am cold" to turn the heat up rather than "Heat up." Another example of invisible interface design is the talking caller ID interfaces that have been showing up on cordless phones. Automatic Number Identification (ANI) has long been a feature of phone services. When displays were added to landline phones, it became possible to see the name and number of the caller if it wasn't blocked. Now, couple the ANI service with a text to speech chip and a display, and a twenty dollar phone can announce who is calling so you don't even have to pick up the phone and look at the screen.

Cars are pretty easy to use today, but they could be easier yet. Purchasing a car that parks itself or avoids collisions is still a little pricey, but cars have all the smarts today such that with a little program my car can help me to schedule a service appointment with the dealer when it knows an oil change is due and I have planned a driving vacation. The car might know that I want the car checked and brought up to standards.

The appointment management system in Outlook does a great job of building dynamic drop down lists that keep places I regularly have meetings, so it is point and click rather than retype. Appointments can be moved or their duration adjusted with simple drag and drop or stretch motions My bookmarks for navigator are well organized and allow rapid access to a variety of different places. I have to manually copy these bookmark sets to the various machines I use, but that is not too much work.

These kinds of software improvements will continue to make use easier, more intuitive, more invisible, and more automatic.

Most laptops and cell phones have wireless networking capability that will recognize and connect to a variety of networks automatically once set up. With a little more work shared printers and drives can be identified and used. While it requires a manual shifting of the profile, this is a simple single click. At home, the wireless network has a hub with a smart modem built in to allow connection to the external network. The in-home network is made up of the wireless hub, and a series of PC and laptops. Drives can be shared for common storage or to allow the various machines to use specialized devices attached to one of the machines – such as a DVD burner or a scanner. Some input devices such as mice and keyboards that make use of Bluetooth can be shared among the machines. Other devices such as cameras and scanners are not shared as easily. Newer TV's and other display devices can use media sharing capabilities. As time goes by, all of these connections and configurations will get easier and allow ever more ad hoc configurations. Checking into a hotel room in the future should allow me to use the display device existing in the room to display the presentation on my cell phone and connect to the hotel color printer to make copies of my presentation or a photo I took that day.

Shifting control to agents

There are a variety of different kinds of agents one might describe. We have talked a little bit about robots, spiders and bots that wander the web looking for information. Here we want to talk about primitive and not so primitive agents that act in primitive and not so primitive ways on our behalf. Let's begin with some simple ones. The receiver for my TV system allows me to program "scouts" that look through and filter all the listings giving me advance notice of soccer games or sci-fi movies with robots. What might more advanced agents do?

I have a large number of web pages that have links to remote websites. I also have a few thousand bookmarks that I maintain. I have programmed a little agent that runs itself every week and lets me know when it finds a broken link or obsolete bookmark. The agent also keeps a textual signature of the remote page, its size, and date of last modification. The agent lets me know when links are broken or when it thinks I might want to revisit the page because it thinks there has been a significant change. The system works well enough for me, but some would find it onerous. With time and a little creative energy, it might be made into a truly useful agent.

When I set up an appointment to advise a student, I want my computer to gather the records on that student for me without me having to ask. It can also fill out the electronic forms that will be required. I do a number of tasks that have a required set of operations where the activities vary very little. In many personal computer programs, there are macro facilities that allow a whole series of operations to be recorded so they can be completed automatically. It is not hard to imagine a larger scale macro system that might be instructed to "watch" and record a pattern of activities such as would be required in advising a student, or preparing a syllabus such that it could intelligently feedback the actions and variables in such a way that I could say "yes, that is what I want done when this condition arises." Over a couple years, you could correct your agent to make it adjust to more complicated situations.

As another "simple" example, consider a question answering agent. Here's how it works. As a teacher, students frequently ask questions. Increasingly, these come as email to a special account. Year after year, I answer those questions. Some are new, other are very predictable. Some are fun to answer, some are tedious. Imagine that we build an agent that operates for a year recording the question, the answer, the course, the student, the week of the term, the degree status of the student, etc. Wait a year, and crunch all of the data into a large database. Start looking at the questions and the context – experience of the student, time of the term, particular course, etc. If a question and context is found that appears to match an existing question or questions, provide the student with the answer immediately and ask if the answer helps to answer the question. If yes, update the database with the new data. If not, pass the question on to me for response and then record all of the data. By this point in my career, I'd be willing to bet that students would be getting immediate answers from my agent instantaneously, and I would be doing less drudge work. Of course, if I made the system work, which I have not yet been able to do, within six months someone would come along and make a half dozen improvements which would create a yet better agent.

This last example is important in a couple ways. The agent I have described is not technologically sophisticated. It is little more than a complicated database lookup and some natural language processing – of the simple kind. The agent begins in a form that is not intrusive, but is useless. Only after this agent observes the simple process a few hundred times does it begin to gain confidence that it can perform meaningful work on my behalf.

Potentially Dramatic Changes

Trying to assess where the most dramatic changes will take place and what they will be is always risky. Even simple trajectories can be twisted and turned by viral applications that seem to come out of nowhere.

I have selected four places where I am fairly confident things will change. The choice was pretty easy – follow the money. In 1981, I contributed a piece to a monograph on Telecommunications and Higher Education. The article was entitled “Telecommunications and the Learning Society”.[72] The basic premise of the piece was a challenge to the belief of several authors, notably Robert Hutchins that we were moving toward a learning society.[73] The article asked how we spend time, and how we spent money, both personal and national. The conclusion to the first question about time was that we were spending all of our newly found time engaged with media – television and radio. That suggested we might be tending toward a leisure society. The question about money was very clear in 1980, and it is even more clear today. That segment of our life that absorbs the most of our money is health care. In that respect, the trend was to a fountain of youth society. Finally, the paper addressed the new forms that education was taking, from Sesame Street to National Geographic to community colleges and non-traditional systems of higher education. TV is not quite the “vast wasteland” suggested by Newton Minnow[74]; it did then and does more so now provide a variety of information perspectives of varying levels of quality. Thus, three areas of our life that are likely to see the impact of the digital information revolution are medicine, education, and leisure. I have added a fourth area, the activity to which we devote the most of our waking time – work.

Medicine

There will be more to say about the social aspects of medical treatments in the next chapter. Here we will focus on the delivery of personal care and the impact of digital information. In the 1980’s, I worked with a group of physicians on structuring medical records using SGML, the predecessor of XML. Dr. Sean McLinden and some of his colleagues were exploring how a schema might be structured using the basic concept of “episodes of care.” The basic idea was that treatment for an ailment might be considered an episode of care. Any given episode might contain other episodes of care – e.g. referral by a generalist to a specialist for some advice, or varying attempts to treat a condition such as high blood pressure. Episodes could be concurrent, sequenced, recursive. In each episode, there

[72] Spring, Michael (1981) *Telecommunications and the Learning Society*. In **Telecommunications and Higher Education**, Institute for Higher Education, University of Pittsburgh, June, 1981, pp. 1-28.
[73] Hutchins, Robert (1968) The Learning Society, New York: Mentor Books.
[74] Minow, Newton (1961) “Television and the Public Interest”, delivered 9 May 1961, National Association of Broadcasters, Washington, DC. Retrieved May 2011 from the American rhetoric website: http://www.americanrhetoric.com/speeches/newtonminow.htm

would be certain common elements and certain optional elements. While other formats are emerging for storing medical records, the notion of using a defined language to capture the events such as is provided by XML is bound to lead to a common format for medical records. Several competing efforts are currently moving forward including the Veterans Administration and a standards groups known as Health Level Seven (HL7) that began many years ago with development of standard formats for transmission of records from various health care devices – such as heart monitors. It will take years to work through this process. With the standards in place, easier input and reporting mechanisms will be required, but they will come. When all that is in place, it will be possible for an individual to carry their entire medical record on a storage device likely to be the size of a credit card. Or it may be the case that information directories on the network are such that the record can be stored centrally and downloaded on demand. The financial impact of a consolidated standardized system for electronic medical records will have a profound impact on decreasing the cost associated with the creation, storage, and retrieval of those records. It will also vastly increase the accuracy of those records. I participate today in a program for electronic management of my health records at the University of Pittsburgh. Communication with my physician has been significantly improved. The results of my annual physical, or any tests I might take, are in my mailbox with commentary before I get home from the doctor. Prescriptions move to my pharmacist before I leave the doctor's office. Questions I might have for an internist who is very busy are answered within 24 hours. Visits to check my blood pressure are eliminated by use of a $20 device at home and an email update to my physician. (And of course, the blood pressure monitor will commonly communicate the results without human intervention – such devices have existed for several years for athletes. It is just a matter of connecting all the dots for all of us.)

But this is only be the beginning. IBM's investment in the Watson project – the computer program that competed on Jeopardy – has as one of its major pragmatic goals the development of a capability to manage large volumes of information and find the answers to questions posed in natural language. Medicine is frequently suggested as one of the applications envisioned for this work. This agent would better enable physicians to manage the enormous amount of new knowledge being developed in the medical field. Consider some of the other agents that might be developed in the medical domain. Medicine provides an interesting environment for agents in the extensive development of the digital infrastructure that can support those agents.

Physicians regularly interact with data about patients. In the case where a physician is dealing with a referral, he or she has to make a decision about whether to trust test results in the patients record.

Increasingly, these tests can have metadata associated with them indicating who did the test, what the machine was, when it was last serviced, what the conditions were, etc. A physician could assess this data, but this is a tedious and time-consuming task. One might imagine a surrogate for the test being assessed by an agent working for the doctor to determine if the test results met the doctors standards for this patient's situation. The agent would have two important components. First, it would build a database of decisions by the doctor that would inform its decision about the validity of existing text results. (A doctor might accept a test related to one area of diagnosis but require new tests in another.) The second component would be a physician friendly interface that would show the tests that exist and the agent's recommendations. It might be a simple green/yellow/red coding of tests with individual physician dependent annotations about key factors behind the recommendation. (The agent would learn the key factors for each individual physician in accepting test or ordering new tests.) The system would have to be designed to have a slow learning curve and it might be important for reward systems to be developed to induce physicians to train and use the agents. Given that the test surrogates are a natural byproduct of the evolutions of digital records, the agents would not be particularly hard to develop. If the social institutions discover that billions of dollars could be saved, it will not be hard to find a way to induce physicians to use them.

Doctors often need to refer patients to specialists. This process involves creating a match between the patient's needs and the skills and recommendations of other physicians. This is a natural environment for a directory of endorsed certificates. In this case, each physician will have a surrogate certificate. We use the term certificate here in the formal sense of signed digital certificates that are issued by a trusted authority. This provides an assurance that we can trust that a physician who is listed as having certain specialties does indeed have them. The next parts though are the most important. First, each physician might be endorsed by other physicians with some indication of the strength of the endorsement. This would not be unlike the recommendation system currently used on systems like LinkedIn.com. The importance difference is that in this system, the identity and qualifications of each endorsement can be verified. The second aspect of the certification might include some follow-up assessment of each referral. Again, these evaluations would be vetted and certified by some trusted authority, but otherwise they would be similar to what students do for faculty on www.ratemyprofessors.com. Such a directory of verified certificates would allow physicians to ask an agent to find someone for "X." In this case the agent would know the physician's criteria for doing the query. I want someone who is endorsed by at least two physicians I trust and who will likely be able to see the patient in the next two weeks and who deals well with older patients. In this way, a physician could ask for a list of doctors meeting certain criteria who had been referred by other doctors known to the doctor making the query.

The key idea here is that we have an environment of digital surrogates that can be processed by an agent which has a set of rules that can be tuned based on the particular wishes of a physician. The agent will have a simplified interface which will present the findings of a significant information processing task to the user. If the agents save time, improve results, and are easy to use, they will grow in use just like modern spell checkers.

Education

Next to the cost of medical care, education is probably the largest single cost we face as individuals and a nation. Access to media – telephone, television, radio, gaming and the internet is absorbing a significantly larger portion of our discretionary income, but the total cost of education, for what some see as a diminishing return, is still second only to medical care.

I have strong personal feelings about education. Basically, I believe that educational systems, particularly in the early years supplement parental education. My children were provided with enriched environments from the day they came home from the hospital. The mobiles hanging over the crib was carefully selected to provide the right level of visual stimulation for infants. From music, to toys, to books, to TV, to video games and computer access, our home was an environment with maximum positive stimulation. Car rides were times for game playing and storytelling. When my sons arrived at kindergarten, they had better reading, writing, and arithmetic skills than their classmates. I don't recall the precise details, but each year their skill level differentials as measured by standard instruments declined. I do not blame the school system for that regression. We had an opportunity on a 24 hour basis to tailor learning experiences for our sons. They were not primarily intellectual. Whether it was swimming, building model boats to race in the stream, finding our way through the city following signs, reading books that talked, we were focused on making learning and doing in all three domains – cognitive, affective, and psychomotor fun and challenging. If parents don't participate, enjoy learning, teach their children, and support and augment the organized system, children will be at a disadvantage.

Beyond the parental responsibility, I have strong opinions about the role of education in our social system. I see two basic roles. The first is to impart a body of knowledge and a set of skills, basic or advanced. The second is to socialize the next generation providing guidance in understanding the role they will play in the workplace. This second goal is controversial, but from a distance it is just all too clear. Many have written about the fact that today's schools are still socializing children for yesterday's workplace. I believe this. Both of my children were strongly encouraged to participate in athletics, which I believe provides the best socialization experience related to teamwork and decision making in a rapidly changing context.

While the elementary and secondary schools provided basic knowledge and skills, and they did a good job, I believe that education was effectively supplemented not only by athletics, but by liberal doses of video games to develop problem solving and visual acuity skills. When they arrived at college, I gave them the same advice I give any parent who asks for advice about higher education. Specifically, this is a unique opportunity provided to children by loving parents to enable them to do three things. First, it will be the most focused opportunity to learn things that you don't need to learn. That is, a course in anthropology or art history may not contribute to a particular job placement, but at college you are paying for the privilege to broaden your horizons in ways only the extremely privileged of a century ago could afford. Second, you are being provided with a protected environment in which you can learn to live independently. Dorm assistants aren't parents, and college is not living alone, but it is an opportunity to understand how to live outside the coddled protection of the home. It is a time for parties, football games, and just lazing around the student center or the city as much as it is a time for classes and learning. Third, and I fear this is the greatest failure of higher education, it is a time to hone your basic communication skills – reading, writing, and speaking – and to refine your analytic and problem solving skills. Note that all of these skills have been taught prior to the college experience, but it is college where you have the opportunity to practice and refine these skills.

Reading, wRiting, aRithmetic and Reckoning

There is little doubt that our lives are dependent upon some basic verbal and analytic skills. Reading and writing are basic communication skills of the literary tradition. Sam Deep, a leading author and consultant in leadership development once commented that there are four communication skills, writing, reading, speaking, and listening and the extent to which we use them are inversely proportional to the amount of time spent teaching the skills. His point was to emphasize the fact that few of us have adequate skills in listening, a sad but true fact. Despite the amount of time spent helping to develop writing skills, they atrophy in most adults. We need to continue to work to develop better communication skills. The environment is changing and the tools available for communication are more diverse, but the basic goals of helping those we are responsible for educating has not changed. We need to help them formulate messages that communicate better and to be able to decode the message and understand what was intended. Books, essays, emails, Facebook pages, twits and videos can all communicate a message, or fail to do so.

Besides communication skills, math skills are important as a means for abstraction and calculation. Obviously, math is much more than arithmetic. The one request I made of my children as they went to college, was to take as much math and logic as they could stomach. One ended up with a philosophy major with minors in math, chemistry, and computer science. The other gave up after calculus and focused on business courses with a strong aversion to the analytic courses in economics and statistics he had to take. Mathematics involves the study of algebra, calculus, geometry, combinatorics and logic. In computer and information science, the less well known and studied fields of mathematics tend to be the more important – i.e. combinatorics or discrete math, logic, and geometry/topology or more specifically graph theory. Again, it is important to not lose the big picture in the minutiae. Mathematics provides tool for reckoning about and representing our environment. To the three R's, I would like to suggest a fourth, Reckoning.

Jeannette Wing,[75] former director of the National Science Foundation and Chair of Computer Science at CMU has called for attention to computational thinking. She argues that computational thinking is a skill everyone needs. It involves problem solving and system design. An important tenet of her argument is that computer science and the other design sciences have developed formal ways of solving problems and designing things that can be taught to people and be used in all human activity. For example, deciding what line to choose in a supermarket involves performance modeling. Modeling graphs, operating on formal structures, building in redundancy, and other techniques can be taught in ways that will allow for better application of the techniques to all sorts issues.

Our educational system will also need to develop techniques to help students come to grips with the power of immediacy in the same way it helps us to become conversant with literacy. The Beowulf of the era of immediacy may well be Gordon Bell's Life Bits[76] project. This research effort developed a comprehensive personal collection that tells the story of

[75] Wing, Jeannette (2006) Computational Thinking Communications of the ACM 49(3), 33-35.
see also Fletcher, George & Lu, James (2009) Human Computing Skills: Rethinking the K-12 experience. Communications of the ACM; 52(2), p23-25.
[76] Bell, Gordon (2011) MyLifeBits. Retrieved May 2011 from the Microsoft website: http://research.microsoft.com/en-us/projects/mylifebits/

one man's life and efforts to record and manage the information store. At some point in the future, a child will be born whose whole life experience will be captured and digitized so as to form an intimate memory supplement far beyond anything that Vannevar Bush might have imagined for his elite scientists. As time goes on, children will master techniques for capturing, organizing, mining, and sharing that personal experience. Some will try and fail and their communications via the integrated technology will be noisy, uninformative, and boring. Others will achieve a simplicity, clarity, elegance and intensity that will allow the receiver to experience the world and its structure in a way they might never have been able to before. I might imagine that a future reader of this book might exclaim that this presentation is a horrible waste. Why did I choose the literary form when I had access to the new digital communication medium.

Primary Education

In 1968, George Leonard[77] wrote a delightful little book about education in the future. Even 35 years later, it still holds up a vision of what could be. On "Visiting Day, 2001 AD", parents toured a school where students interact individually and collectively with dynamic interactive displays. There is "Ongoing Brain-wave Analysis (OBA) that guides Computer Assisted Dialogs (CAD). The children interact with the system using speech recognition technology, keyboards, and gestures. Some of the other visions go even further out in time.

Leonard gets several things right, even if he imagined a much greater change in 30 years than has occurred in 40 years. First, he imagines a world where human computer interaction in primary education is very personal and interactive. We are closing in on this today, but even in the US we are still some distance away from one laptop per child. He is also right on target about the fact that we need a model of the learner that helps to determine what needs to be done next. Individually prescribed instruction (IPI) has been an area of active research and development in the education field for about 40 years. To make IPI work, we need to map the hierarchy of skills that lead to some desired target. This gives us a roadmap of the skills that need to be taught. Now we need some information about the student and their performance. Do they need concrete examples? Do they respond better to auditory or written presentations? Did they get the initial examples correct? Do they need more practice? With a model of the learner and a model of the subject matter, it is possible to adapt the instruction to the individual student. Most recently, with the mass of data

[77] Leonard, George (1968) Education and Ecstasy. New York, Dell Publishing Delta Books.

we can collect from thousands of users on the web we can begin to develop aggregate data models of different kinds of learners. As we move to the future, teachers will continue to develop as coaches and managers and more and more of the direct instruction will move to computer systems with their models of the subjects and the students.

As you might imagine, given a highly personal system connected to a network, it might not be necessary to have a centralized school. That is unless both parents are working and we take seriously the belief that another role for school, besides cognitive skills, is socialization. The options here become more complex than I can manage. In a little bit, I am going to talk about telecommuting to work. That means parents might again be at home and able to provide the needed level of supervision. Just as with remote work today, school work can be intimately monitored remotely and coaches dispatched when there is an issue with progress. Then, places for social interaction and group work can be smaller and used when necessary. My crystal ball is foggy as to the direction we will move. The best guess I have is that our continued investment and desire for physical infrastructure will stymie the development of a suitable digital infrastructure.

Higher education

In **Education and Ecstasy**, Leonard suggested that higher education would disappear and that people would engage in specialized learning as needed throughout their life. I am ambivalent about this view. I do think, institutions of higher education will continue to exist for several reasons. One is the opportunity they provide for socialization. Another relates to the fact that for many of our institutions, research is as important an activity as teaching. Universities also provide a preservation function where scholars spend time reflecting on what we know and how we know it. Finally, many young people are not ready for the workplace and families would like to provide an opportunity to develop a more well rounded understanding of the world.

In some areas, higher education will be most appropriate at a time in our lives when we are shifting jobs and need new skills. Imagining that a first professional degree in computer or information science will serve well for a lifetime of work is foolish. Just as some business schools require some amount of time in the workforce before admitting people to their graduate programs. Thus, it is not surprising that there is a significant increase in the number of online professional graduate education programs. They are constructed in a way that allows someone who is working full time complete advanced study more easily.

Today, higher education institutions are experimenting with online courses made possible by the high speed network and the ability to easily capture and combine video and audio in an integrated stream. To be successful, we need a different physical infrastructure complimented by a significantly more sophisticated technical infrastructure. E-business is not simply about the use of technology, but about improving the bottom line via technology. The same can be said about e-education. It should be a better product produced more cost-effectively and delivered at lower cost. Clearly, education is a bit business that is currently housed in an atom format, so we need to repackage the product. If you install course management software that decreases faculty productivity, you are not engaged in good e-business. So, it should be the case that effective online education is better, easier, faster, and more efficient for both faculty and students. It should open new markets, or dramatically improve customer satisfaction. What is it that we are selling? It would seem that there are three things.

First is the exposition of knowledge and skills in a combined literary and oral format. Books contain knowledge in a much more precise and comprehensive form than lectures. So why don't students just read books? The answer is that it is possible to rephrase precise knowledge in a form that is more understandable for those first coming to grips with it. Faculty spend a lifetime trying to come to grips with the knowledge in their field and are able to tailor the important aspects of the information so that it can be more easily understood. Of course, and importantly, not all faculty are good at this. The best are.

Second, we are providing the learner with an opportunity to demonstrate what they have learned and get feedback from someone qualified to assess their performance. This second aspect may even be more difficult to do well than the exposition of knowledge. A great faculty member structures experiences that allows the student to demonstrate or not demonstrate their mastery of the content. When students comment on a course saying "the assignments were meaningless", there is a good indication of failure in the instructional process. For me, the ideal assignment is one in which the student is capable of assessing their own performance in process. (As an example, students in one of my introductory courses have to build a website in which all the pages are "xhtml compliant" and meet the W3C accessibility guidelines. There are programs that will examine pages and provide a detailed report of where they are non-compliant. In this way, a student can compose and check a set of pages iteratively until they achieve compliance.) Unfortunately, not all assignments, especially more advanced integrative assignments can be checked by programs and require a personal professional assessment.

Third, in higher education, we provide a certification that an individual has a certain level of skills. It is not as direct a certification as is provided by primary and secondary education and it often relies on institutional historical reputation. My undergraduate college served to produce many individuals who went on to medical and law schools. While I was never able to confirm it, the rumor was that the top recommendation was pretty much a shoe-in to the best professional schools. It is very clear to me that I can place graduates of our programs in various organizations based on a simple call. Two organizations in particular employ more than thirty of our graduates and managers are aware of the skills they brought with them. They are also aware of the fact that I will not recommend an inferior student.

Consider for example a multiple choice test. In class, you give the exam, score it, give it back, and discuss it in class. If you are doing it online, the test can be different for each student, students can be given immediate feedback, branching can allow a check to see if the question may have been confusing or whether the student might really understand the concept. Immediately after the test is administered, review material can be suggested based on the analysis of the answers. Wouldn't that be something?

So how can we make e-education better than p-education? I would suggest several things. First, make the courses world class. Offer only a few courses on each subject. Design the course to capture the entire market. The content, presentation, services, and the experience should all be first rate and designed to replace all equivalent courses offered by any other institution. There will be tremendous resistance to such a model, but consider a model where the institutions of higher education subscribe or syndicate the best faculty across institution and supplement the course with good faculty focused on providing additional services and feedback. Wouldn’t it be exciting if the offerings of institutions were cut a hundred fold while the students in each of those offerings were increased a hundredfold. Each institution would offer its world class signature courses and students would have the benefit of a combined education across institutions that was unparalleled by the offering in any other form from a single institution. Imagine what the dynamics would be if you could have a class taught by the best faculty and serviced by the best junior faculty and PhD students with small group discussions among the 1000's of enrolled students going on 24 hours a day. These courses would use computer programs and student records to construct personalized materials for each student enrolled. Course associates would work to continually adjusting to the particular needs and learning difficulties of the enrolled students.

Questions addressed to the instructor would be computer mediated. The question, context and answers are stored in a database. As I described earlier the questions and answers would be processed such that the system might eventually answer many routine questions. Those that could not would be triaged up the ladder to assistants, then junior faculty, and then the lead faculty member. From the students point of view, attention becomes immediate but is still personal and individual. Finally, meta analysis of the data after a period of time might suggest revisions to the material

Time would have to be spent to develop effective social systems for the courses. Students might like to talk with colleagues before or after a lecture. They might want to set up study groups. Students in physical proximity might want to meet face to face. The college experience in the context of these great courses could be maintained, but it will take work. If 1000 students enrolled at $1000/seat for the best class on X in the world, each offering would bring in a million dollars. If the basic course structure required an investment of a half million dollars, and delivery and support of the course cost $750,000/ offering, the investment would be at a breakeven point after two offerings.

Other courses might be packaged and sold much like is the case in elementary and secondary education. Unlike the world class courses described above, there are others that might be depersonalized and offered as basic skill courses – again supported by a technologically enhanced human infrastructure. Building blocks are course components that are worth building for reuse. (The argument might be somewhat reminiscent of the move to consolidate statistics courses years ago.) Students might be required to complete basic skill sets before moving to more advanced courses. Consider topics such as "how to properly cite references in a paper", "the assessment of statistical measures used in a research paper", "how to make notes on a book", “measures of central tendency and deviation."

E-education is in our future, but we have not yet taken the time to plan an articulate set of goals, or made the investment to build the kind of infrastructure that makes this next generation of quality educational experiences a reality

The Learning Society

What is described above suggests a system of education that supports learners throughout their lifetime. Whether that moves us closer to a learning society is still not clear to me. The vision that Hutchins put forward brings to mind an environment where people want to learn for the sake of learning. I imagine a nation of renaissance men and women. There is some evidence that in our television of abundance, more and more people reference watching a cooking show, or a history presentation, or some other media presentation focused on knowledge and skills purely for reasons of learning.

Universities are making use of the internet to present the results of their finest minds for viewing by the general public without financial requirements. Open Culture is a website that provides access to such materials. It serves as a single point of access for cultural and educational media. Their website states their mission "*is to centralize this content, curate it, and give you access to this high quality content whenever and wherever you want it.*" [78]

There are hundreds of "nuggets" – pieces of courses that can be mined from the knowledge already well formed in faculty – that abound in institutions of higher education. I believe that just about everyone who has taught and done research for 10 or more years finds themselves at a cocktail party where for some reason they are motivated to explain what they know best to one of the guest's that shows a genuine interest. They wax eloquently for about a half hour and make clear something the guest could never have understood by reading for days. There is, on average, one nugget per senior faculty member at a large research university. Granted, some faculty will be barren, but there will be others who have four or five. At a place like the University of Pittsburgh, there are likely some 3000 nuggets that could be mined. At 30 minutes a piece that is 1500 hours of stimulating and provocative content. I would further be willing to bet that it would be relatively cheap to mine, and that at least 150 hours could be combined to form some new degree for people from 50-70 who want to know a little about all the aspects of our world from first rate minds that can explain it to an educated person. I am willing to predict that some enterprising university will find a creative way to manage this knowledge for a profit during the next few decades.

Entertainment/Leisure

It is difficult to get good consistent data on time use over the last century. Historical data can be obtained from studies by Szalai[79] and Robinson and Godbey.[80] Good sources of more recent data include the American Time Use Survey collected by the Bureau of Labor Statistics.[81] Related specifically to media, good sources include the PEW Internet and

[78] 350 Free Online Courses from Top Universities (2011) Retrieved May 2011 from the openculture website: http://www.openculture.com/freeonlinecourses

[79] Szalai, Alexander (1973) The use of time: Daily activities of urban and suburban populations in twelve countries. The Hague, Mouton.

[80] Robinson, John & Godbey, Geoffrey (1997) Time for life: the surprising ways Americans use their time. University Park, Pa. : Pennsylvania State University Press.

[81] Bureau of Labor Statitics (2010) American Time Use Survey. the Bureau of Labor Statistics website: http://www.bls.gov/tus/

American Life studies[82] and those of the Henry J Kaiser Family Foundation.[83] Combining these various sources allows a couple trends to be seen. Over the last forty years, men spent less time at the workplace and more time doing housework. Women spent less time at house work and more time at the workplace. For both men and women, free time increased from 1965 to 2005 by about 5 hours – to 38.5(women) and 41.8 (men) respectively. All of that time was absorbed by increased viewing of television and video – rising to about 15 hours for women and 17 hours for men. Hobbies and reading papers were the big losers and PC and internet usage grows to about 1.5 hours. Keep in mind these numbers reflect time usage for all adults. Looking at children, the Kaiser study [84]shows some interesting trends for the period 1999-2009. First, for 8-18 year olds, overall daily exposure to media of all types increased from 7:29 to 10:45, but because many young people increasingly multitask, e.g. use the computer and watch TV at the same time, total daily media use increased from 6:19 to 7:38. This is indicative of an increase in multitasking from 16% of the time to 29%. Computer usage tripled from 27 minutes in 1999 to an hour and a half. Video games also nearly tripled from 26 minutes to about an hour and 15 minutes. These numbers probably don't yet reflect the impact of increased network speeds for home broadband, and more importantly smart phones. (The first release of the iPhone was mid 2007.) One of the impacts of the iPhone is more use of the media during times such as bus riding, waiting in traffic, or other situations which historically did not allow access to media or the internet.

Basically, we know that leisure time will likely continue to slowly increase and that media use will continue to absorb more and more of that time, particularly for the younger generation. In addition, there will be more concurrent use of media in the future. It will become a more common for people to search the internet, text message, or browse Facebook while watching TV or listening to the radio. Using streaming video for phone calls will bring grandparents to use the internet, especially when it provides regular access to their grandchildren. The Kaiser report (8-18 year olds) shows correlations between heavy media use and lower grades. These kids tend to be more bored and more often sad. Whether media involvement is an escape mechanism for these youth or a causative factor is an open

[82] Pew Research Center (2011) Internet & American Life. Retrieved May 2011 from Pew Internet website: http://www.pewinternet.org/

[83] Rideout, Victoria, Foehr,Ulla & Roberts Donald (2010) Generation M2: Media in the Lives of 8- to 18-Year-Olds. Retrieved May 2011 from Kaiser Family Foundation website: http://www.kff.org/entmedia/mh012010pkg.cfm

[84] Rideout, Victoria, Foehr,Ulla & Roberts Donald (2010) GENERATION M2:Media in the Lives of 8- to 18-Year-Olds, Kaiser Family Foundation, p 2.

question at this point in time. There does not seem to be any evidence that we are spending more or less time in socialization or communication. They appear to have stayed constant over the last several years. What does seem to be true is that we are spending more time on computer mediated communications. While texting and tweeting, and Youtubing have all grown, the most dramatic growth related to the social networking sites. Not everyone uses or likes Facebook, but it is an unmistakable force. Consider Figure 31[85] which shows the percent of internet users reached by Google Facebook and Yahoo. Facebook is gaining slowly, but look at the number of minutes visitors stay on the site. Facebook visitors stay almost three times as long.

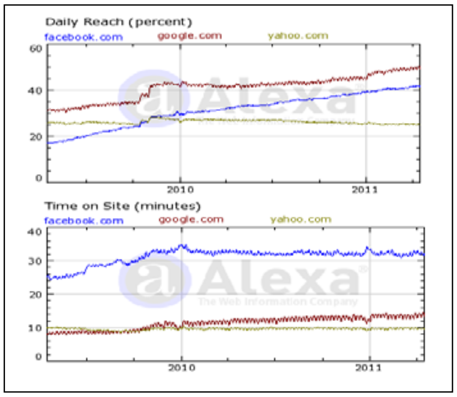


Figure 31: Page Views and Time on Site

Many vendors of hardware and software are banking on a "three screen" strategy, which is the belief that people will continue to use mobile phones, computers, likely touch screen laptops, and television. I believe this is true. I also think it is likely that the screens will be less tightly tied to the devices. That is, I should be able to use a large display to view TV and computer simultaneously. My HD TV's have a resolution of 1920 x 1080. This is currently the highest resolution for the 16x9 format. For the 4x3 format QXGA defines a resolution of 2048x1536. The highest resolution currently define is QSXGA 2560x2048. The two monitors I have on my home PC provide me with a resolution of 2560x1024. I could add two more monitors, if I purchased the appropriate card and get full QSXGA. This would easily allow me to watch two HD videos simultaneously with space left over for a couple video feeds or other information. It is important to note that the standards allow the development of hardware to manage the multiple displays as a single virtual display. An alternative approach would be to have a set of displays where some would be combined and others would stay separate.

Beyond increasing resolution and display from multiple sources simultaneously, people are looking toward 3D TV. Obviously, 3D TV with and without special equipment is already here. There are still issues and standards aren't yet set, but it is pretty clear that we will have some limited 3D sources for selected events. It would be pretty easy to do 3D for the evening news, but who cares. Sporting events will prove a big draw and several events are already being recorded in 3D. Will 3D become an

[85] Site information captured from alexa.com for Google, Yahoo and Facebook on April 18, 2011.

immersion technology – like the StarTrek Holodeck? We already have a Holodeck like experience emerging with the game consoles. There is a vibrant community committed to the development of worlds with avatars in the form of Second Life.[86] Finally, street view in Google Maps provides a way into the real world via our desktop with full control of the perspective. Experiments have successfully mapped real time video into both Google Maps and Microsoft Bing Maps.[87] So, our ability to interact with and observe our world from our home or cell phone will continue to improve. In some situations we will want total control as is provided by the various mapping services. Such control will probably be extended to select environments. In other cases, very expensive production values will be applied to control the presentation of important events where we can see somewhat more than we can in the current environment.

Work

Thinking about the future of work in the age of digital information brings science fiction type visions. The vision begins with telephone operators. As a child, I lived on a college campus and when I went to see my father at work, I passed by the telephone operator who manually switched calls to the offices of the administrators. Fifteen years later, I attended a college where incoming calls were still switched manually by patch cord. In both cases, the operator was placed in a very public location in the main administration building. The story of the telephone operator is that if phone calls were still connected manually, it would require the world's population to connect all the calls being made today. (Actually, the story is old enough that I suspect today there would not be enough people to switch all the calls made today.)

Vanishing Jobs

Where possible, we continue to develop machines and programs to take over those work activities that can be done better by machine. Telephone operators, cow milking, produce picking and hundreds of other occupations are losing ground to machines and programs. For other occupations, systems are being put in place that obviate the need for workers. Bank tellers have given way to systems – ATM's and online banking – where consumers do the work for themselves replacing the need for workers dedicated to interfacing to specialized systems. Grocery clerks were replaced early, but today, many retail outlets have self checkout options. Online trouble shooting systems have replaced many product

[86] http://secondlife.com/
[87] http://searchengineland.com/bing-demos-streetside-with-flickr-photos-live-video-bing-sky-35994

specialists, or at least the systems have been centralized. One might imagine that it will not be too long before the demand for postal workers begins to decline. Already, electronically delivered statements are beginning to become the norm and customers are beginning to be charged if they want paper statements.

New Opportunities

On the other hand, there are some jobs that seem to be relatively immune to the digital information revolution. These include all the personal service providers such as hair stylists, trainers and physical therapists, advisers such as financial consultants and lawyers, child care workers, teachers and medical personnel who provide individualized services. It would seem that police and firefighters will continue in secure jobs as will trades people from carpenters to electricians to painters to plumbers. Also included here are cleaning people, gardeners and other such workers.

There are also new jobs that are emerging. Machines are not yet able to take care of themselves. The machines are getting better at being self diagnostic, but we still need auto mechanics to fix a fuel pump or change the oil. ATM machines require humans to fill them with cash and diagnose problems. Computer and televisions are getting closer to being disposable, but technicians are needed to fix them when they break. The medical arena has hundreds of devices that require sophisticated care, from pharmacy robots, to various forms of scanners, to blood pressure monitors and weight scales. The science fiction vision of this class of jobs is not that machines become our masters, but that we are so dependent upon them that we need to expend great amounts of energy to take care of them.

Two other occupations that appear to be growing and will likely continue so are deliverers and greeters. We used to go shopping and bring what we purchased home in our car. Today, increasingly we are purchasing small items individually over the network and having them directly delivered to our home. The growth of UPS and the other delivery services are indicative of this trend. Online experiences work hard to maximize customer satisfaction. This is one of the major changes in e-business which we will discuss in that chapter. It is expected that both online and offline businesses will provide personalized customer service. You can chat online with most major retailers today and similar services are increasingly present in physical stores – the Wal-Mart greeter and the super store clerks who today always seem to have enough time to take you to the row where what you are looking for is located. Home Depot offers mini classes on how to do it yourself, all in the context of providing more personalized services.

The Deprofessionalization of Everyone

Through the 1960s and 1970s, there was a vibrant discussion of profession among sociologists. A big part of the discussion had to do with the huge body of technical information that professionals had to manage. Basically, professionals are, from a sociological perspective, individuals who make use of knowledge and skills not accessible to the general public. Professionals provide services of import to individuals and are accorded a high status because they act on behalf of the clients for the greater good. Some argued that as more professions developed skills and bodies of knowledge that could not be accessed by the general public, more occupations would be accorded the status of professions. Others made a counter argument that said as the knowledge and skills became available to the general public, it would cause a trend toward deprofessionalization.[88] Will the availability of digital information and smarter programs make us all competent jail house lawyers or reasonably competent medical practitioners? I think it is too early to make any judgment on this matter. Clearly, the best in the various professions not only have significant knowledge, but significant wisdom gained from the repeated application of that knowledge.

Outsourcing, Off-shoring, Contract Work and Telecommuting

Beyond the observations made above about various occupations that may be waxing or waning, I think a couple of trends that we are seeing today will continue. First, telecommuting is increasingly well supported by the infrastructure. For information workers, the need to do the work in a particular time frame and space will decrease and organizations will find ways to accommodate highly valued workers who can be more productive living where they want and working from there. Also, as organizations focus on their core competencies, the low cost of transactions on the internet make it cheaper to outsource the work. From the perspective of an individual worker, networks may make it possible to make an excellent living as a contract worker. Significant adjustments will have to be made in health care and other benefits to establish ways for this class of worker to exist and survive. Of course, outsourcing and contract work, when they are to overseas organizations result in off-shoring. The state of the information business is such that increasingly, the services that are outsourced can be outsourced to other countries with little more difficulty than they can be outsourced down the street.

[88] Haug, Marie (1973) Deprofessionalization: An Alternate Hypothesis for the Future, The Sociological Review Monograph 20 (1973): 195.

The Balkanization and Federation of Personal Information

Each of us may be looked at as a source of information – personal information about who we are, what we do, what we think, etc. Increasingly this information is leaking out of the places where we have kept it securely or insecurely. Joseph Luft and Harry Ingham proposed a tool for thinking about interpersonal relationships.[89] Named for the authors, the Johari window poses four quadrants in which we might place behavior and motivation. We use it as a tool for thinking about personal information:

Johari Window for Personal Information		Known to self	
		yes	no
Known to others	yes	open	blind
	no	hidden	unknown

On the web, we consciously expose some personal information, and make an effort to hide other information. What is most interesting is that organizations may be able to uncover personal information about me that is not known to me. Every time Amazon makes guesses about my interests based on gift purchases I have made for others, I am reminded that sometimes they think they know something personal about me, but they have instead failed to separate gift and personal purchases. Some information about me gets federated. For example, the policy statements of banks and online businesses indicate that while they keep your information private, they share it with a variety of the organizations in their federation. Other information is Balkanized. Medical and dental records take some effort to move as do hospital records.

Ownership and shared information

This leads one to interesting notions of the ownership of information and the location of information. Perhaps the most interesting information in this category is medical information. Consider a person's medical record. Who owns this record? Is it the physician who created it? Is it the patient it is about? The matter may be further complicated by considering the components of a medical record. Who owns the following:

- An x-ray of my lung.
- A reading of the x-ray
- A diagnosis based on the reading

[89] Luft, Joseph, Of Human Interaction. Palo Alto, California: National Press Books, 1969.

"Order" of information

The notion of an x-ray and a reading of the x-ray suggests that information might be thought of as having an order. Some information, such as an x-ray or the fact that it rained in Pittsburgh today might be considered primary or first order information. Reading an x-ray might be considered derivative or second order information. Information such as "it rained on 110 days in Pittsburgh during 2010 is also second order. Collecting this information for several years, one could derive a piece of third order information – "for the last twenty 20 years, the average number of rainy days in Pittsburgh is 28%. Who owns derivative information? For example, the fact that a person has a certain height or weight might be defined as first order information. The fact that a person is "overweight" is obtained by a function of height and weight in accord with some algorithm. Is the information that a person is "overweight" their information or does it belong to the person who applied the algorithm?

High grade versus low grade

Some information that describes me explicitly – my height, my medical condition – might be considered high-grade information about me. Low grade information might include:

- What books I purchase from amazon.com
- When I am logged onto the internet
- What Microsoft applications I use, or what I ask for help about

Do I have the same right of ownership of high and low grade information. Indeed how is the ownership established? When low grade information is anonymous and aggregated, who owns it? Generally, I do not control this kind of information about me. When is information explicitly or implicitly transferred? As with the shared ownership issue, this category again asks who actually owns what portion of the information.

Information Doping

We are increasingly able to define formal operations on information. While we know how to copy and delete information, it is still difficult to define identity or transformation operations. For example, as derivative information is developed, can we formally describe the transformation and measure the changes in information caused by the transformation.

From an end-user perspective, the operational problems with information include such things as understanding the ownership of information, the rights of various individuals or organizations to disseminate that information, or how to trace the flow of the information. If we are concerned about keeping private information private – how do I secure the information about me that for one reason or another I divulge to others? How do I insure it is not reused without my permission? Is there a difference between selling my email address to others, telling others I am looking for a mortgage, or telling others I work at a University?

What is the role of public-private key encryption in controlling access to “jointly-owned” information – e.g. a physician can keep information about me, but it may only be released when accompanied by both my and the physician’s private keys. How does this impact medical research? How does this impact emergency care? What happens to my information when I cease to exist? How does my privacy impact the public right to be safe?

Can information be doped? Explosives are tagged or doped with trace chemicals that allow the origin of the explosive to be traced in the event that it is used in a crime. Like some compression, the doping process is asymmetric. It is low cost to dope an explosive and high cost to trace the tags. Like compression, the basic idea is to keep the costs of the most frequent operation low and allow the costs of the less frequent operation to grow. Thus, when a file is infrequently compressed and very frequently decompressed, the cost of the compression can be high while the cost of decompression is kept low. How might doping or watermarking be used trace illicit use or dissemination of information?

Chapter X: Social Perspectives

Social Networks and Social Capital

We are social animals and regularly participate in a variety of social networks. Over the last century, sociologists have been studying social networks. Jacob Moreno[90] in Europe and William Lloyd Warner[91] at Harvard took various approaches to the study of how people interact in social networks with the use of ethnographic field methods in the study of industrial communities. A social network is a set of relationships between individuals who are members of a bounded social group. The term "bounded" indicates that there is some mechanism for defining who is and is not a member of the group. Membership can be as tightly defined as family members or formal membership in an organization or as loosely defined as a shared interest.

The history of social network research is rich and varied. Milgram's study of message passing is frequently cited as an early example of social psychology research in this area.[92] For Milgram, the "small world problem" was finding an explanation for why it was possible to send a communication from one person in a large social group to another person in a very "few" steps. Work on social capital emerged during this period as well. (Social capital is added to the array of kinds of capital we have at our disposal – financial, physical, human, and intellectual. Social capital is most simply envisioned as the trust we have in others. If I trust you, I might let you use my car, or borrow money without any formal collateral.) The concept was formulated by James Coleman[93] and popularized by Robert

---

[90] Moreno, Jacob (1951) Sociometry, Experimental Method and the Science of Society. An Approach to a New Political Orientation. New York: Beacon House.
[91] Warner, William & Lunt Paul (1941) The Social Life of a Modern Community. New Haven, CT: Yale University.
[92] Milgram, Stanley (1967) The Small World Problem. Psychology Today, 1(1), pp. 60 – 67
[93] Coleman, James (1988) Social Capital in the Creation of Human Capital, American Journal of Sociology (Supplement) Volume 94, S95-S120

Putnam through several articles and his book Making Democracy Work: Civic Traditions in Modern Italy.[94] Putnam's research attempts to explain how the strength of social networks is critical to the fabric of society. The development of socially oriented web sites on the Web extends how we are able to interact and creates some new form of social interaction. In 1996, I had the opportunity to attend an invitational conference at Xerox PARC during which Putnam asked a very simple and direct question. Can the Web provide a mechanism for restoring the social networks that have been on the decline for the last several decades[95]? The answer fifteen years later is a resounding yes. In this chapter we look at how some important social interactions are changing, and how new kinds are being created. Before turning to more explicit social networks, we look at a phenomenon of growing importance related to incidental social activity on the Web

Wisdom of the Crowd

There is an old bit of folk wisdom that suggests that sidewalks not be built between buildings in a complex like a college campus until after the users of the environment have worn paths in the lawns. That will tell you where the paths need to be placed. This demonstrates a form of "wisdom of the crowds" that is being looked at seriously on the Web today. Besides using log information to determine how people use a website and where they want to go, we can use information about where they go to inform many types of decision making. The importance of an academic article has long been rated in part by the number of citations. The more researchers refer to a given article, the more important it is. If academics working independently all find an article to be of import, it must be. One of the many factors used by Google to give you the best results from your search first is just this phenomenon. Called Page Rank, named after one of the founders of Google – Larry Page, the score is generated by a rather complex recursive formula. Simply put, a page that has links coming from many other pages is likely more important than a page with few incoming links. Adding a little more complexity, we assess the importance of each page pointing to the page in question and rate the page more important if the pages pointing to it are themselves important. So when you type a search like "baseball legends" which just identified 140,000 pages that matched the

---

[94] Putnam, Robert (1993) Making Democracy Work: Civic Traditions in Modern Italy. Princeton: Princeton University Press.
See also Putnam, Robert (1996) The Strange Disappearance of Civic America. American Prospect. Volume 24, Winter 1996, pp 34 - 48.

[95] Moline, Judi (1997) Conference Report: First Annual Leveraging Cyberspace Conference. Journal of Research of the National Institute of Standards and Technology, 102(3), pp 371-374. Retrieved May 2011 from the NIST website: http://nvl.nist.gov/pub/nistpubs/jres/102/3/j23mol.pdf

terms and wonder how Google picked the first ten it displayed, a big part of the answer, but far from the complete story, is how other people felt about the various pages Google found. As you search the web, write blogs, upload and tag photos, rate videos, bookmark pages, you are engaged in social activity that is being recorded and used by various agencies on the web to make decisions based on the wisdom of the crowds.

New Communities and Interactions

Through the first half of the Twentieth century, most American lived either in the city or the country. With the prosperity following World War II, the population of the suburbs exploded. If you look at houses in residential city neighborhoods, you will note the large porches facing the street that established the local community. Block parties were, and in city neighborhoods, are still common. As the population moved to the suburbs, back yards, decks and patios became the context for socialization, maybe around a pool or barbeque. In the 80's and 90's the party moved inside and frequently became limited to the family – in aptly named "family rooms." As media devices became cheaper and began to populate multiple rooms, individual rooms began to see more personal activity. We seldom even watched TV together any more. Today, even in those solitary environments, we are beginning to revive our social interactions. Communities of interest are growing around our old classmates, professional colleagues, distant family members using a variety of mechanisms and media.

Families

One of the first things to happen on the Web for families was the development of photo sharing sites, the most famous is Flickr, which was created in 2004 and claims about 20 million users. Families can share photos through any of the websites they may have created as part of their Internet service. I recall how amazed people were in the late 1990's when I would post a page of thumbnail images on my website linked to full size photos. For most people, simpler processes were needed. Webshots, one of the earliest sites, was created in 1995 and claims more than 30 million users. Snapfish, created in 2000, may be the largest claiming 70 million. Wikipedia lists more than 30 major sites.[96] These sites allow grandparents to see photos of grandchildren and friends to share vacations and special events.

[96] Wikipedia (2011) List of photo sharing websites. Retrieved May 2011 from the Wikipedia website http://en.wikipedia.org/wiki/List_of_photo_sharing_websites

Facebook has emerged as a significant presence. It has become an integrated environment for posting images and carrying on conversations. It is likely there will be a consolidation with photo sharing sites and social networking sites. As security features allow careful restriction of information access, families will likely make greater use of such sites to record important and confidential family information. It is worth noting one of the newer sites, Windows Live Photo, which was created in 2008, claims 56 million users. Windows Live Photo is a part of the Windows Live effort which seeks to integrate a series of Windows services available on the network. Thin clients on Windows boxes will provide access to secured cloud based storage and services. Live services already exist to establish identity and manage relationships. There are also services to manage parental controls as well as services for managing groups.

Skype has established a major presence in computer based video chat, and after a number of attempts has established a reliable video conferencing system that is available with their Premium Package. Microsoft also has Windows Live services that include video chats and, for a cost, Live Meetings that provides more extensive sharing services than would be required to support family interactions. Google talk does not currently support group video calls, but does have other integrated services. At the current time, all these video calling systems are not interoperable and require all the parties agree in advance to set the system up. With time, better integration will be achieved. Further, agents will add alerts and other kinds of sensors that might make it easier to monitor a senior in their home, or know when they have not taken their medication, or turned on lights in the evening. There will be a delicate balance between privacy and appropriate levels of monitoring. It is not hard to envision an environment just a few years down the road that would link a camera in one home to a screen in another based on pre-agreed conditions. For example, if I am at home, and the motion sensor in another home detects an individual in the kitchen, turn on the PC and camera in the kitchen and establish a video call to the screen in my home. In some such way, a natural level of agreed to interaction might be established to provide a sense of security to a parent or senior relative. Similar setups might be created for children arriving home from school before the parents return from work. It would establish an automated check-in that would be minimally disruptive to the work environment and would provide a natural environment in which a child could update their parent on changing plans

Social Groups

One of the findings arising from the social network research of Robert Putnam is that overall social commitment was tied to active participation of members of the community in social clubs. Affectionately referred to as the animal clubs, social organizations like the Elks, Moose, Eagles, Lions, Masons, Rotary, and Kiwanis have declined in recent years. Similarly, Putnam found a decline in bowling leagues. These are places which provide for cross organizational social networks and establish a commitment to civic life. While the web has not increased participation in these kinds of organizations, direct civic engagement and political activism appears to have been reinvigorated by the Web. The engagement in social causes such as the movement for reform in the Middle East in 2011 has been spurred by the availability of information on the Web. More telling, the Web social networks appear to have supported the protests by allowing information about activities to be spread instantaneously. Politicians are now able to gather data from large groups of citizens in near real time. How the crowd mentality, with its rather fickle nature will merge with the kind of representative democracy that is our history is not yet clear, but it is clear that the communication networks are being opened.

We go online more and more frequently to find health information. The Pew Internet group found that the number of Americans who looked for health information online increased from 25% in 2000 to 61% in 2009.[97] In addition to information seeking, peer to peer support groups are arising spontaneously in various areas. Studies have examined family and social networks and their impact on coping behavior for issues such as bereavement.[98]

Work is also being done on restricted social networks to support individuals who have been found to benefit from some form of cognitive behavioral or group therapy. Some of our own work at the University of Pittsburgh[99] has shown significant benefits to individuals with schizophrenia, caregivers of individuals with traumatic brain injury, and individuals with various forms of cancer. There are many people who fail to benefit from these forms of treatment because they are reluctant to meet face to face, or have difficulty scheduling time for such treatments. The

[97] Fox, Susannah & Jones, Sydney (2009) The Social Life of Health Information. Retrieved May 2011 from the PEW website: http://www.pewinternet.org/~/media//Files/Reports/2009/PIP_Health_2009.pdf.

[98] Breen, Lauren & O'Connor, Moira (2011) Family and social networks after bereavement: experiences of support, change and isolation. Journal of Family Therapy. 33(1), pp 98–120. Retrieved May 2011 from the Wiley website: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-6427.2010.00495.x/pdf

[99] Steel, Jennifer et. al. (2011) Randomized controlled trial of a collaborative care intervention to manage cancer-related symptoms: Lessons learned. Clinical Trials.

Rotondi, Armando et.al. (2010). Web-Based Psychoeducational Intervention for Persons With Schizophrenia and Their Supporters: One-Year Outcomes. Psychiatric Services. 61(11) pp. 1099-1105.

Rotondi, Armando., Sinkule, Jennifer & Spring, Michael (2005) An Interactive Web-Based Intervention for Persons with TBI and their families: Use and Evaluation by Female Significant Others. Journal of Head Trauma Rehabilitation. 20(2), pp 173-185.

web makes it possible for these people to interact with an appropriate amount of exposure and on a schedule that meets their time constraints. In addition, some groups who might be connected are too widely dispersed to cost effectively provide the treatments face to face. For example, caregiver support groups for military personnel who have suffered traumatic brain injury need to provide access for individuals who are spread across the country.

The kinds of social interactions made possible on the Web continue to increase in diversity. Some sense of the diversity of these groups can be found on meetup.com. It combines the best aspects of digital and physical social networks by providing a digital directory of local groups and their activity. When accessed in April of 2011, there were more than seven million members in almost 80,000 local groups involved in about 250,000 monthly meetups. Looking for meetups near my home, I found an Earth Day meeting, a French language group, a meeting on revitalizing vacant lots, a yoga group, a pro-choice group, a running event, and of course several Steeler and Penguin groups – and that was just groups meeting today and tomorrow.

Another interesting development is the emergence of uncompensated altruistic contributions in the context of collaboration. While I post many of my materials to the web, it is an individual, not a social effort. Contributions to public sites such as Youtube are similar. (One of the new interesting forms of contributions is Photsynth which is a Microsoft effort that allow people to take sets of photographs which can be merged and displayed in Bing Maps.) But as a social construction activity, nothing compares with Wikipedia. Begun as a companion project to a Web encyclopedia project free to the public but written by experts (Nupedia), Wikipedia was started in January 2001. Wikipedia quickly eclipsed Nupedia and hit a period of accelerated grown around 2004. Hundreds of scholarly publications have explored the Wikipedia phenomenon. The plethora of early pieces criticizing the quality of some articles has given way to trying to understand how the system functions and how it might be duplicated in other environments.[100] What is clear is that people have an interest in contributing and working together and derive positive results from the interaction. It is a safe bet that other efforts such as Wikipedia will emerge in the future to meet our social needs to contribute.

[100] Lam, Shyong & Riedl, John (2011) The Past, Present, and Future of Wikipedia, IEEE Computer, 44(3) pp. 87-90.

Organizations

In the next chapter we will address the impact of digital information on business, both online and offline. In this chapter, we reflect a little on the social life of individuals in organizations. One of the significant impacts is the tethering of individuals to the workplace. Earlier we talked about the possibilities for telecommuting on the positive side. On the other hand, more and more employees are on 24 hour call, or at least many have extended work hours by phone and email. An increasing number of communications are sent from always connected devices. Mail arrives in the inbox on a regular basis 24 hours a day, and even individuals who are not paid to think about work 24 hours a day are often checking on their boss or the status of some work effort from their phone while they commute or from a home PC. In what I consider an unhealthy aspect of this phenomenon, an employee's regeneration time can be corrupted by a demanding supervisor.

It will be possible, with a significant amount of work, to begin to manage the knowledge in our organizations. Much of the information in organizations exists in the minds of its employees. The goal of knowledge management is to capture that information so it can be shared in the organization. Such information has been shared in social networks over the years. Copier repair persons know what the manual says about fixing certain machines, but they also know that some repair people have found workarounds and tricks that are not recorded in the manuals. Historically, this information was shared through peer to peer exchanges that occurred after work sharing a beer or through informal work exchanges. Similarly, at the professional level there are mentoring exchanges for young faculty and young researchers. There are no good manuals for writing a course syllabus. Some people have significant insight, gained from writing proposals and reviewing them, that can be shared with other researchers.

The organizational goal here is three-fold. First, organizations have to find ways to maximize the opportunities for informal sharing – e.g. we devoted a portion of a seminar to brainstorming ways to create digital social networks that simulated the kinds of exchanges that sometime occur in after work exchanges over a beer. More specifically, the informal exchange in a bar is not the reason for the meeting, but a bi-product of it. How can an enjoyable social experience be provided digitally that will maximize the likelihood of such exchanges. Second, organizations have to work to eliminate the barriers to sharing. Not all sales people are created equal – some are much better than others. However, raising the bar by making mediocre sales people superior may not be viewed as being in the best interests of the superior sales person who might view sharing as losing their edge in the organization. In some ways, this is simply a matter of changing the reward structure such that the collective effort is rewarded as well as the individual effort. While easy to say, it is not as easy to modify deeply embedded incentive plans.

The third and most interesting aspect of an effort to manage the knowledge in an organization is building systems to capture and refine it. I developed a collaborative authoring system in the mid 1990's designed to capture as much activity as possible.[101] The times that users were connected were recorded along with their actions. The system would visualize all of the activities by all of the users on all of the document components. Ad-hoc hypertext structures about the documents made it possible to track the number of accesses, number of minutes open, number of comments, etc. These data were used to assess user attitude toward the document, construct ballots, inform editing, as discussed in the section on the social periphery on page 141. While the system was built to reduce the cost and improve the time to develop collaborative documents, it provided the beginnings of a knowledge management system. By capturing ALL of the exchanges between individuals in the work environment it is possible to capture some, or maybe most, of the informal and tacit knowledge that people have. With improved natural language processing, and using contextual metadata – who was involved in the communication, when did it take place, what was going on in the organization at the time, etc. – it should be possible to locate and process high-grade information that might be formalized into explicit knowledge. Right now, few organizations have a system in place to capture all of this data, and fewer still have mechanisms in place to try to distill the information in an effort to find important knowledge.

Dangers of Social Networking

Social networking is still in early stages of development. One indication of this is the rapid evolution of interfaces as new strategies and techniques are tried, replaced, augmented and dropped. With time the flaws and shortcomings of early efforts will be eliminated. The two most important areas in which changes will occur are security and customs. On the security side, techniques for defining our circle of friends and exposure of information will emerge. They will be simple to understand and implement. Regards customs, we will learn what to say and not say in this new social environment. Stories of stalking and cyber-bullying appear in the news with alarming frequency. Children have always taunted each other, but the visibility and communications reach of bullying on line is

[101] The CASCADE system – Computer Aided Support for Collaborative Authoring and Document Editing. See http://www.sis.pitt.edu/~CASCADE

unprecedented. Further, our sites follow us. They can be embarrassing at one level, but they can also be used in job interviews, legal proceedings, etc. Stories of these events and personal experiences will slowly help each of us to understand what we should and shouldn't say. A colleague once advised me that if I had something good to say, I should do it in writing and if I had something bad to say, I should deliver it in person orally. What we say at a cocktail party or in a bar is different from what we want to say at a meeting when newspaper reporters are present. It will take time, but we will learn. I recently asked one of my former graduate students if she had gotten married – noting a new last name on her Facebook page. She said no, she was just using a different last name for security reasons. Similarly, I read recently that the best photo for your Facebook page is that of a bald man who is deceased! It may not be particularly helpful for old bald men, but useful for many others.

There are some dangers associated with social networking. As parents of children or as mature adults, we need to pay attention. The first danger is one of time commitment. Television viewing, or video game playing, or texting are not dangers in isolation. It is excessive use that is a danger. We might imagine that each of us has a certain amount of time at our disposal. We can use it for social networking, online shopping, or video gaming. The time we commit might be reasonable – especially if it begins to replace anti-social activities to which we have become addicted. The time commitment could be defined as intrusive if it prevents us from doing some of the things we would normally be expected to do. Personally, I get frustrated when household chores, or professional commitments take second place to social networking. Finally, we could define the time commitment as disruptive if it becomes our highest priority at the expense of many things we are committed to such as work and family.

The other danger of social networking is the overexposure of private information. As stated earlier, some people use pseudonyms to reduce their exposure. Others use avatars or funny photos as personal images. Yet others seem bent on exposing every aspect of their daily life without constraint. In the physical world, we constrain our comments in public in a variety of ways. We don't talk to an interviewer in the same way we talk with our parents or children. After a half decade, individuals have come to understand that Facebook pages are not only viewed by those for whom they are intended, but by a number of other people. There are two ways to deal with the issue of exposure. First is to very tightly control access to the information we post. Thus a social networking site that is restricted to a tightly controlled group – e.g. family members – allows us the freedom to say what we want knowing it is being viewed by only a select few. But for many, this kind of social networking site is not exactly what they envisioned. Many people want to use these sites to gather together friends and casual acquaintances. In this case, it is important that people understand that those who can view what we post include four groups – those for whom we intended it (friends at college), others for whom it was not intended but who are welcome (our cousins), those for whom it was not intended but who represent no active threat (parents, police, potential employers), and those for whom it was not intended and who represent active threats (identity thieves, stalkers, etc.). We need to learn to write and post conservatively or in highly controlled groups.

Nation States

The Internet and the new forms of digital communication are having enormous impacts on nation states. The uprisings in the Middle East in 2011 are one powerful example. The use of YouTube and twitter to organize protesters in near real time was unprecedented. Over the last decade, politicians have begun to use social media as a means of organizing grassroots activity. We have also begun to see the emergence of the threat of cyber terrorism. These topics are briefly discussed.

Politics

It is likely that digital information and social networks will continue to have an impact on politics. Beyond the ability to poll and share information, politicians have the ability to use the network as a means to raise funds. Stories of internet fund raising abound and go back as far as the campaign of John McCain in 2000. In 2004, ActBlue was established to raise funds for democratic causes. As of April 2011, it had raised more than 185 million dollars. Contributors are able to target their contributions to particular causes – as of April 25th, 2011 more than 38,000 donors had contributed more than \$850,000 to “stop the Republican war on working families.” The various campaigns have specific targets, mechanisms, and contribution goals. Howard Dean made use of Meetup, text messaging, and online videos to engage and activate people. Increasingly, mass customization techniques are being used to target messages to different constituents and to solicit funds for particular efforts. Barack Obama’s election in 2008 included significant use of the Web and text messaging to get the crowd out and keep people informed.

Political blogs abound as do Facebook pages and YouTube appearances. Candidates now have the ability to quickly reach out to where the crowds are – and that is the social networking sites. Some fear that there will be targeted campaigns using email and text messaging that will provide information to people in somewhat different forms based on an assessment of what people want to hear. Historical polling feedback to politicians is reaching a whole new level of immediacy both directly and through news outlets that gather the data and then broadcast it via traditional channels to the masses.

Like businesses, federal, state and local governments are increasingly using the Web to provide faster services – such as e-filing. They are also using the web to increase transparency and provide interested constituents access to information about government. Electronic balloting has grown more slowly and with great concern that the systems might be susceptible to tampering and fraud. The problem is compounded by the infrastructure expense in the transition and by the fact that the responsibility for voting is widely distributed. It is not hard to imagine a time when electronic balloting using personal certificates will provide a secure, convenient, and speedy way to vote. Prototype systems exist in many organizations where a user logs into an intranet and then submits an electronic ballot. The system assures the anonymity of the vote, and prevents people from voting more than once. It is true that the system needs to keep track of who has voted in order to make this work, but the same is true in face-to-face voting environments.

Crime and Surveillance

Digital crime will replicate existing crimes and take some new forms – most of which we have already seen. Since statistics have been gathered, fraud and failure to deliver services have been at the top of the list of crimes.[102] The majority the various forms of fraud relate to online auctions, either misrepresentation of the product for sale, or failure to deliver the item purchased. Significant efforts have been undertaken to fix the systems, and improvements have been made, Indeed, the most recent annual report of the Internet Crime Complaint Center (IC3) indicates the rise of scams where information requests appear to be coming from the FBI and an increase in identity theft

> *The most common victim complaints in 2010 were non-delivery of payment/merchandise, scams impersonating the FBI (hereafter "FBI-related scams") and identity theft. Victims of these crimes reported losing hundreds of millions of dollars.*[103]

Crime goes where the money or property is easy to steal and that is increasingly the Internet. One of the scariest parts is the massive nature of the digital information stores that can be accessed. Stealing credit card

[102] Internet Crime Complaint Center,(2010) 2010 Internet Crime Report) Retrieved May 2011 from the IC3website: http://www.ic3.gov/media/annualreports.aspx
[103] Internet Crime Complaint Center, (2010) 2010 Internet Crime Report, page 4.

numbers has been going on for years – copying the number from a receipt or using a handheld scanner when processing a transaction. While there are numerous wait staff at restaurants who could steal lots of numbers in aggregate, the ability to get millions of names and numbers from an insecure database is just too tempting. Phishing expeditions, where an email asking for login information is sent to a million users, don't have to have a high success rate to garner information on thousands of users. The amount of information that can be obtained and the ease with which it can be processed makes this form of cyber crime extremely attractive.

Another form of cybercrime that is difficult to assess is bribery. The data is closely held because the source is organizations that operate on the basis of customer trust. Indeed, the Department of the Treasury's Financial Crime Enforcement Network requires that banks and other financial institutions report various classes of suspicious financial transactions, but also make it illegal to disclose such reports, which are called SARs or Suspicious Activity Reports. Basically, the bribery involves asking for payment from an organization to not disclose a demonstrated vulnerability. Rather than taking customer data, the criminal would demonstrate that they were able to access the data and asks for money in exchange for not disclosing the fact. The other form of bribery supposedly involves networks of bots that the criminal controls. Unless criminal demands are met, a distributed denial of service attack (DDOS) is threatened to make the organizations site unavailable for customer access. DDOS attacks have been executed for a number of reasons and it is believed they are increasingly being used for organized crime purposes.

The number of cameras being installed in cities worldwide continues to grow. Both London and Chicago claim more than 10,000 cameras on streets and public spaces controlled by the police. Public CCTV cameras are also being used for traffic monitoring. Pennsylvania began to install cameras on major state road in 1992 and now manages more than 620 statewide that are also accessible to the public. In addition, closed circuit television cameras (CCTV) are widely used in banks, airports, hotels, schools, stores, bars and famously casinos. Indeed for the last several years I have kept four cameras attached to the internet active, two of them in my office. It is estimated that where cities have public surveillance cameras that number in the tens of thousands, the total number in large cities number in the millions. The police systems are increasingly using powerful face recognition software that can identify persons of interest. Software is also becoming available to track individuals across camera interfaces. That is, an individual who is identified can be tracked as he leaves one cameras field of view and moves into another. There is little evidence that cameras deter crime, particularly in public areas, but there seem to be a growing body of evidence that such cameras are playing a role in apprehension and enforcement. (There is growing evidence that red light cameras are playing a positive deterrence role.[104])

Camera surveillance is only one form of tracking that is possible with the current technology. As our phone of choice has shifted from the wired to wireless network, it has become a device that can be tracked. Initially, the FAA required that the network providers make it possible to locate a cell phone for police and fire agencies. The whole 911 system was predicated on the basis of being able to get to someone in an emergency. Initially, cell phones thwarted this need. Today, with the advent of smart phones that not only communicate over the cell phone network, but contain global positioning system chips that are capable of reporting highly accurate location information. In 2011, it was revealed that the Google Android phones and the Apple iPhones were collecting tracking information from phones.[105] Google and Apple both claimed that the information was anonymous and used only to improve services. While this may be the case, thousands of smart phone users are downloading free applications that request access to storage and location information and the vast majority of users don't examine whether personal information is being logged or sold by the application developer.

Other stories of GPS and network tracking abound. There are a variety of tracking security products – from laptops to cars and commercial vehicles. They even offer a bracelet that can help you keep track of your loved one – my family used it with a senior uncle with early onset Alzheimer's who had a tendency to walk off and get lost. People who are of concern to law enforcement can also be tracked. A number of states have begun program that track parolees – particularly sex offenders. Some have suggested that it could be of use in tracking individuals guilty of domestic violence. Currently, the cost of such tracking is high, and it is not clear that it is effective. For the most part, this is a matter of developing better software to monitor the inputs. Beyond tracking, smart device data can be added to the information. My local electric provider is offering a rebate to customers who allow tracking and reporting of electricity usage. An insurance company is offering lower rates – up to 30% – to customers who allow their driving habits to be profiled. It makes a record over thirty days

---

[104] Hu, Wen; McCartt, Anne. and Teoh, Eric (2011). Effects of red light camera enforcement on fatal crashes in large US cities. Arlington, VA: Insurance Institute for Highway Safety. Retrieved May 2011 from the IIHS website: http://www.iihs.org/research/topics/pdf/r1151.pdf

[105] Newman, Jared (2011) Phone Location Tracking: Google Defends, Apple Stays Mum. PCWorld Apr 24, 2011. Retrieved May 2011 from the PCWorld website: http://www.pcworld.com/article/226156/phone_location_tracking_google_defends_apple_stays_mum.html

of when you drive, how many miles you drive, and how frequently you make sudden stops.

Warfare

There have already been numerous changes in the complexion of the battlefield. What most differentiates the US military from others it the quality of the digital information they command. The global positioning system was developed by the Department of Defense in the 1970's and became operational in 1994. It allows precise navigation of weapons systems and military personnel. While the program was been put on hold in 2009, the Future Combat Systems program provided some idea of what the battlefield of the future would be like. Many of the weapons and surveillance systems we see today arose in part out of that program. Included are the unmanned aerial drone and various robotic ground vehicles. But perhaps most impressive were the visions put forward for the Future Force Warrior. The ground soldier would now have the kind of heads up display used by pilots providing informed friend or foe identification and support for different vision modes as well as an integrated communication system. Uniforms would also be advanced with better personal armor and science fiction like exoskeletons. Each soldier would also be kept comfortable with cooling and heating systems embedded in their clothing with extensive sensor systems providing feedback on their health.

On the cyberwarfare front, there are significant concerns about three things. First, the Future Combat Systems are based on extensive data gathering and communication. If systems can't gather data from the battle field or if commanders can't communicate action plans to their units, these advanced systems deteriorate in value. Second, information about the military and about the US in general is stored in computers and that information is attacked on a regular basis. Spies don't have to gather data in the field – it is collected by us and stored in computers where it can be accessed from a distance. The issue here that most concerns the military is that like cybercrime in general, attacks in cyberspace do not rely on the same projection of force that is the hallmark of the physical battle field. Referred to as asymmetric warfare, a small force of highly skilled hackers can do severe damage to information systems from a distance. The third concern has to do with the protection of national assets – or critical infrastructures. The first two concerns are now the responsibility of a new unified command in the US military. The US Cyber Command is responsible for the security of military networks. It is also charged with the task of ensuring US/Allied control of cyberspace and denying same to adversaries. Some people are concerned that this says the military should have plans in space to control the Internet. I believe that ultimately it does mean that. One of the huge differences with digital information is that it is not easily bound by physical boundaries. At the same time, technically, protection of the non military network, the national assets, is the responsibility of the Department of Homeland Security.

Historically, systems were over built to survive extreme use conditions. In the U.S., for the transportation system, this is the Thanksgiving holiday. For the phone system, it is Mother's Day. Most of the critical infrastructures lived on specialized networks. For example, until recently, electrical distribution systems operated on networks separated from the Internet. Increasingly, power and other critical infrastructures make use of the Internet. This means that the underlying networks need to have rerouting capabilities to insure that expected system failures are minimized and localized. The network must have survivability in the context of catastrophes such as fires, earthquakes and typhoons or war. Beyond insuring the security of the base network, which is not a simple task, the critical infrastructures are now potentially exposed to attack through the network. Responsibility for securing the critical infrastructures is the responsibility of Homeland Security, but the military shares responsibility. As of 2011, seventeen critical infrastructures have been defined.

- Agriculture and Food
- Banking and Finance
- Chemical
- Commercial Facilities
- Commercial Nuclear Reactors, Materials and Waste
- Dams
- Defense Industrial Base
- Drinking Water and Water Treatment Systems
- Emergency Services
- Energy
- Government Facilities
- Information Technology
- National Monuments and Icons
- Postal and Shipping
- Public Health and Healthcare
- Telecommunications
- Transportation Systems

The Homeland Security Presidential Directive 7 charges DHS with coordinating the protection of critical infrastructure protection by establishing a framework to identify, prioritize, and protect the critical infrastructure. Thus, various federal agencies – Treasury for banking and finance, Health and Human Services for public health and healthcare – share responsibility for carrying out plans in the critical infrastructure sectors.

Chapter XI: Impact on Business

Commerce versus Business

Commerce is the "the exchange or buying and selling of commodities of goods." Interestingly, the primary definition of commerce is "social intercourse; the interchange of ideas." Post-industrial societies are increasingly involved in the exchange, buying, and selling of information and ideas. Thus, electronic commerce or e-commerce is the electronic exchange of information and the support for the exchange of other commodities. Some e-commerce companies – eBay, E*TRADE, etc.—are engaged in the exchange of information. If we accept that eBay as an auction house is only involved in the buy and sell decisions, not the exchange of the physical entities, it is a good example of pure e-commerce. Another good example would be E*TRADE. However, for most businesses, some portion of the exchange is physical. Thus, amazon.com has a significant commitment to physical storage and shipping, as well as to managing returns. Amazon.com may be viewed as equally divided between e-commerce (the order) and p-commerce (the delivery). There is every reason to believe that more and more pure e-commerce businesses will blossom in the coming years.

In contrast to the term e-commerce, the term e-business focuses on how the business operates. An e-business may or may not be involved in e-commerce. The notion here would be that the business itself, the organization, is heavily invested in the electronic realm. An e-business is an enterprise working via electronic means. Any number of facets of a business might be carried forth electronically. Consider as simple examples the following business activities that might be carried forth in electronic form.

- recruiting of new employees
- marketing, advertising and public relations
- customer support and education
- meetings and information resource sharing among employees

- training of employees
- intelligence gathering for strategic and tactical planning
- distributed inventory control functions
- payroll and benefits management

It is relatively easy to imagine how any of these activities, if appropriate to a particular business intent, might benefit from instantiation in an electronic form. They will all have an impact on the bottom line. There are three points here:

1. An e-business may, of course, be engaged in pure e-commerce. The focus on e-business is not meant to suggest that e-commerce is excluded. Indeed it is likely that e-businesses will be engaged in e-commerce and that e-commerce companies will have a strong commitment to being e-businesses.
2. Not all aspects of a business can, should, or will be electronic. A traditional business can be made more efficient by making use of digital information.
3. Some of the most profitable and rapidly growing aspects of e-business are those that involve the streamlining of back office processes to adopt a more efficient electronic form. Specifically, business to business transactions that enhance supply-chain management, just in time inventory control, and accounts payable and receivable are among the first targets of opportunity for smart business people.

E-business is concerned with the growth and development of new and existing businesses to make better use of digital information. Consider as just one example, supermarkets. Supermarkets were early adopters of point of sale inventory tracking. They also began mass customization to target marketing using coupons printed at the register in response to the items purchased. With ID cards they began customer tracking and profiling.

Key Concepts for E-Business

It is difficult to understand the myriad of announcements and developments that occur every day. Some businesses are too concerned with using the latest technology and not concerned enough with building a better business. Concepts that transcend particular technologies or processes are important tools in understanding what is going on.

Atoms to Bits

One of Nicholas Negroponte's messages in his wonderful little book **Being Digital** has to do with advantages of bits over atoms. Basically, the message is the following—if your business product is really information (bits) and it is being exchanged physically (atoms) rather than digitally, you will need to migrate to the bit form of exchange because the relative costs dramatically favor digital exchange. This has already occurred with music and is now occurring with books. This principle causes us to examine our business carefully to determine if we are indeed using the most appropriate form. The economics of bits over atoms compel those in bit businesses to move from atoms to bits.

The financial sector has dramatically embraced this philosophy. The cost of a bank transaction over the counter is about $1.07 and an ATM transaction is about $.27. In contrast, an internet based transaction is on the order of $.01. When one considers the number of banking and brokerage transactions daily, the numbers are staggering. In addition, online banking is viewed as a customer centric service allowing clients to access the services 24 hours a day. This is despite the fact that the client is doing much of the work! Today, many banks are still heating/cooling and lighting branch offices in high cost environments. While there will continue to be reasons to maintain physical infrastructure, it needs to be carefully justified. Indeed, it might be the case that the compelling reasons for buildings might give way to more customized and personal services delivered at a time and place convenient to the customer – i.e., rather than maintaining branch offices for the few face to face transactions that need to take place, banks might dispatch loan officers to a person's home when the need arises.

Publishers of music, video and text are also clearly on the move. Education and medicine will follow

Inventory to Information

> *"Extraordinary improvements in business-to-business communication have held unit costs in check, in part by greatly speeding up the flow of information. New technologies for supply-chain management and flexible manufacturing imply that businesses can perceive imbalances in inventories at a very early stage--virtually in real time--and can cut production promptly in response to the developing signs of unintended inventory building."*[106]

[106] Greenspan, Alan (2001) Testimony of Chairman Alan Greenspan, Federal Reserve Board's semiannual monetary policy report to the Congress Before the Committee on Financial Services, U.S. House of Representatives, February 28, 2001

Greenspan's observation highlights one of the many important reasons why an e-business replaces inventory with information. Business agility in transitional times is important, but there are more reasons. Inventory represents sunken money that cannot be used for other purposes. It also occupies space, which may be expensive and requires management. It is subject to theft and obsolescence. And finally, it is not easily analyzed.

Organizational Publishing

Publishing related expenditures reflect 6% to 10% of the gross revenues for the Fortune 1000. In some industries, such as securities publishing costs can be higher. From annual reports, to stock prospectuses, to monthly bills, to paychecks, organizations print hundreds of documents that must be distributed and then collected as waste. It appears that 2010 will likely be a turning point in digital delivery of information from banks. Through 2010, incentives were offered by many financial institutions to shift from hard copy to digital delivery. In 2011, we began to see indications that where possible financial institutions would move to make physical documents a cost item.

Besides the fact that the delivery of bits is far cheaper than the delivery of atoms, organizations can't afford to maintain duplicate infrastructures. While bits are close to ten times cheaper than atoms, the human costs to write and edit the information are pretty much the same. If the costs of bit based printing and publishing are half the cost of physical printing and publishing, when both structures are in place, they will represent 9%-15% of gross revenues. While the transition is made, there will be a decrease in productivity. Once the transition is proven to be effective, the old infrastructure must be significantly reduced. Organizations cannot maintain file servers and filing cabinets.

Transaction Costs and Organizations

In 1937 Ronald Coase wrote an article about the nature of the firm.[107] For that article and other work on property rights, he was awarded the Nobel Prize in economics in 1991.

> *First, as a firm gets larger, there may be decreasing returns to the entrepreneur function, that is, the costs of organising additional transactions within the firm may rise. Naturally, a point must be reached where the costs of organising an extra*

[107] Coase, Ronald, (1937) "The Nature of the Firm," Economica 4(16), 386-405. Retrieved May 2011 from the Wiley website: http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0335.1937.tb00002.x/full

> *transaction within the firm are equal to the costs involved in carrying out the transaction in the open market, or, to the costs of organising by another entrepreneur.*[108]

At the risk of oversimplifying Coase's elegant arguments, which are still fun to read today, transactions can take place in an open market, where there is a cost for the transaction. Within an organization, the transactions are coordinated by the business – i.e. there is a coordination cost. When the cost of a transaction goes down, the size of the firm shrinks. More accurately, transaction costs and coordination costs interplay to determine the size of the firm. Transaction costs include search costs, contracting or negotiation costs, and transaction execution costs. Transaction costs today – search, negotiation, and execution – have all decreased dramatically and firms are restructuring themselves to focus on their core competency. Outsourcing is optimization of core expertise in firms that can now transact in the open market.

As the information infrastructure continues to improve, it is likely that many new forms of organizations will evolve with a mixture of open transaction and coordination characteristics. For example, in supply chains, suppliers may be categorized as strategic collaborations – scarce products, strategic cooperation – important products, short term agreements – products for which price reductions can be negotiated, or commodities – products best purchased on the open market. In each case there is a different level of integration that is desired. Given good information systems, suppliers who once viewed an organization as a black box may now be given an appropriate view of the internal details of the organization without compromising private company information.

Customer Centric

E-business focuses on the customer in new ways. While cost is an important part of business, it is far from the only factor in establishing a relationship with a customer. The other major factors are the quality of the product or service and the service provided. Increasingly, businesses of all types are providing customer rewards. Beginning with groceries and drug stores, loyalty rewards programs are appearing everywhere from big box stores to restaurants and hair salons. These loyalty program allow for customer tracking and the provision of better service. Given lower communication costs, there are an increasing number of reminders about upcoming auto maintenance, and scheduled hair cut time. Since the auto

[108] Coase, Ronald, (1937) "The Nature of the Firm," Economica 4(16), p394.

companies track the cars I purchase from them, it's not very surprising when I begin to get more aggressive updates about specials and deals that are coming up right about the time I am ready to purchase again.

Push to Pull

When an e-business really has its act together, it is possible in many cases to begin to make the shift from pushing product to allowing customers to pull the product or service. Historically, Dell has been held up as the best example of a pull oriented e-business. Rather than the push model that involves strong marketing of products, the pull model tries to stay on top of the customers preferences. Because it is a component assembler, Dell is in an ideal position to keep low inventories and to communicate to its suppliers the changes in customer preferences. Dell saw the move to LCD's from monitors early on and responded to the pull by working with vendors to change the product mix being manufactured. While Dell is easy to perceive as pull oriented, other businesses demonstrate the same sensitivity in less dramatic ways. Supermarkets can watch trends in product purchases and expand variety in areas of customer interest and contract offerings in other areas. Pull orientation is easy to see when it is done well, or poorly. Just as Dell tries to insure that it can immediate construct a computer with the components you choose, colleges and universities should seldom have classes that are closed!

RSS feeds have become the mechanism of choice for making the shift from push publishing to pull publishing. They also provide the major mechanism for mass customization of news and information. As of April 2011, syndic8.com listed almost 650,000 feeds. Checking for gardening, Syndic8 provides a list of 47 feeds. The PCWorld site provides access to podcasts and 17 blogs and more than forty product feeds from smart phones to games to operating systems. I can use my browser or mail reader to subscribe to the specific information I want and pull it from the organizations that I want to pay attention to. If I don't want to build my own feed list, sites like feedzilla provide access to individual feeds, personal aggregation, and preaggregated categories – more than 2000 predefined news categories.

Mass Customization

The nineteenth and twentieth centuries saw the development of mass production technology. The current state of manufacturing technology, using computer controls, allows us to begin to realize mass customization. Obviously, customer centric businesses that have a commitment to "pull" will also lead in customization. Whether it's a individually designed personal computer or custom jeans, many businesses make products that are manufactured in such a way that customization is possible.

Levi Strauss & Co. had the capability with computer control to make one-up custom jeans in 1999. The system required that the customer visit a store where their measurements could be taken and sent to the factory. Two weeks later, the customer could return to pick up their jeans. Analysts have suggested all sorts of reasons why the program failed – the time between order and delivery, the lack of a comprehensive customer relations management program, cost, etc. After a half decade out of the mass customization business, the company is returning to mass customization in a new way – called CurveID, it allows individuals to make a decision online based on shape rather than size. An array of companies provide simple forms of mass customization – such as photos printed on tee-shirts and coffee mugs. Others provide small batch production – energy bars, chocolate, and pet food. Still others provide custom jewelry, shirts, and shoes.[109]

Netflix and Amazon provide mass customization in another way. Watching my purchases, Amazon alerts me to sales of similar items that I might want to see. They also use the purchasing patterns of others who match my profile to cross sell and up sell. Unfortunately for Amazon, most of my online purchases are gifts for other people and generally reflect their needs rather than my own. None-the-less, Amazon makes me feel like they pay attention to my personal interests. Netflix may set the standard for mass customization of a service. Using a complex algorithm, Netflix recommends movies I might like to watch. Indeed, in 2006, Netflix offered a one million dollar prize to anyone who could improve the algorithm. In 2009, the prize was awarded to a combined team of competitors. I find Netflix to be amazingly accurate in its predictions even with the complicating condition that it has to predict based on both my and my son's selections. I am not sure how it handles the bifurcated choices of historical documentaries and Japanese anime!

Global Markets

E-Markets -- Adam Smith proposes a theory of economics that suggests that optimal trading takes place when the buyer and seller have perfect knowledge of the market. The size of the market is limited by our ability to obtain knowledge about the market using the available technology. From the bazaars of the mideast to the seaport marketplaces of Amsterdam to the commodities trading floors, markets have expanded to

[109] The website http://mass-customization.info/ provides a reasonable picture of the growing efforts toward mass customization

the limits available. All of these markets were limited by space until the last fifty years.

Today, the marketplace is worldwide. It is possible for the buyer and seller do develop very sophisticated knowledge about their counterparts. The Internet has created a globally sized market where the players have an unprecedented opportunity to gain perfect knowledge of the vast market. The Web has made it possible for each consumer, sans intermediaries, to access the sellers in the market place directly. Businesses can present themselves directly in the market place. Consumers can bypass intermediaries and buy directly. A vast new marketplace is open to businesses, but they must keep in mind that consumers are no longer bound by geographic constraints. I make use of a business in Hong Kong to buy cell phone accessories at 15%-25% of the cost of the same item in local stores.

"Information" and Reengineering

Shoshanna Zuboff, in her book "The Age of the Smart Machine", suggested that the computer is doing for intellectual processes what the steam engine and other power technologies did for manual processes. Power equipment allows us to automate work processes. This restructuring of work is called automation. The application of the computer, with its memory and logic allows us to restructure knowledge work. Parallel to automation, she calls the process information. Where power technology automates, information technology informates. E-business is characterized by informated processes.

Wherever possible, work processes should be examined to determine if a process can be made better, faster, or more efficient by the use of computer technology. In this process, innovators need to be careful not to overspecialize processes as so often happened with the application of power technology. Information technology has the power to integrate work activities returning control of whole processes to individuals. For example, it becomes possible to provide a loan officer with the tools that enable a loan to be approved by a single individual using a computer to perform the specialized subtasks. Thus, what has become a disconnected set of subprocesses in many banks, leaving the person applying for a loan with a sense of alienation, can once again be integrated and streamlined.

Through the early 1990's, there was much attention paid to business process reengineering. Hammer and Champy in their book "Reengineering the Corporation: A Manifesto for Business Revolution" promoted the streamlining of business operations. One way of looking at reengineering from a technical point of view is to see it as adopting practices that are made possible by the deployment of information technology. In this sense, reengineering provides the business principles that guide what Zuboff called information. At the core of Hammer and Champy's approach to process reengineering are several principles. These include:

- replacing assembly line type jobs with case work type jobs
- moving decision making down the chain of command
- converting management positions from authority control to coaching and training
- triage of cases into a range of cases characterized by simplicity/complexity
- reduction of the number of non value added checks and controls

This process can streamline business, but it can also create major employment dislocation. In traditional organizations, there can be a lot of resistance to reengineering. As importantly, the informated systems may be less capable than intended. Basically, the goal is to replace humans with data and algorithms. While stories of the significant improvements made in organization abound, there are also horror stories. The success stories talk about order of magnitude improvements in speed or cost. The down side is that there are high failure rates.[110] One of the basic tenets in using technology to streamline organizations is working to automate low level decision making and processing and move the human role to a more integrative position. It is interesting that the process of automation created a kind of specialization that led to employee alienation in the workplace – i.e. workers did not make a complete product in which they could take pride, but simply played one repetitive role. Information, on the other hand appears to provide an opportunity to move the basic repetitive tasks to computer control and provides an opportunity for the human to play a more integrative and control mode. Known as Blauner's U-curve of technology integration, the theory suggests that applications of technology initially alienate employees, but as high levels of integration are applied, workers find themselves participating in the work.[111]

Business Information as a Commodity

For many businesses, information can be viewed as a secondary business. Successful internet businesses such as Google, Yahoo and Facebook use information about their visitors directly to target advertisements. Increasingly, organizations are providing selected

[110] Bashein, Barbara, Markus, M. Lynne, and Riley, Patricia, (1994) Preconditions for BPR success. Information Systems Management, 11(2), page 7-13.
[111] Hodson, Randy (1996) Dignity in the Workplace Under Participative Management: Alienation and Freedom Revisited. American Sociological Review, 61(5), pp. 719-738.

information about their clients to affiliates who then target market to them. Other business information can also be of value. For example, trends in product sales may be of value. The key is insightful aggregation of business information.

Transforming Business

A business provides goods and/or services to consumers. In a completive environment, businesses try reduce the costs of production so as to sell their product or service more cheaply increasing market share. Alternatively a business may work to differentiate their product or service from those of competitors such that their product is chosen when costs are similar. The basic question that an organization needs to address is how the digital information infrastructure changes the means available to reduce costs or improve quality. Related to these choices it is interesting to consider that if an organization earns a net profit on what it produces of 7%, reducing costs by $1 dollar is equivalent to increasing sales by $13. (Thios may not be intuitive for some. Think of it this way. If I sell $100 worth of product and make a 7% profit, it means that $93 was the actual cost of making product and $7 was the profit. If I divide $93 by $7 I get 13.28. This means that if I sell another $13 worth of product, I will make another dollar profit. I accomplish the same goal by reducing the cost of producing the product for sale by one dollar – one dollar in cost reduction increases my profit the same amount as selling $13 more of product.)

It is likely that one or more of the following may be true:

1. the Internet provides information and communications tools that makes it possible to order the goods and services needed by the business reducing excess inventory, insuring smooth processes by assuring needed stock, and reducing the transaction costs for materials acquisition.
2. the Internet provides a new channel through which products may be marketed. This marketing will either open a new market, allow the retension of an existing market, or improve the relationship with the existing market.
3. the Internet provides an opportunity to restructure the organization in such a way as to reduce the processing costs that support the business.
4. the Internet and the information stores in the organization that have grown historically or that can be developed as a result of 1,2, or 3 above provide an opportunity to develop a new or auxiliary product or service .

Numerous books have appeared on how to transform business activities, both conceptually and technically. They all address three basic tasks: internal management of an organization – enterprise resource management, management of suppliers – supply chain management, activities that pertain to consumers – customer relations management.

The resurgence of the American economy over the last few years is closely associated with the reengineering of the American corporation over the last decade. The streamlining and restructuring of business in line with the opportunities provided by computer and communications technology is less glamorous than e-commerce, but more profitable from a bottom line business perspective. Closely associated with restructuring of business to be more e-business oriented is the development of business to business commerce using digital information. The most dramatic growth in e-business over the next several years is projected to be in the area of business to business commerce.

Organizational Efficiency

Consider as simple examples the following activities which exist in some form in most organizations.

- recruiting of new employees
- meetings and information resource sharing among employees
- training of employees
- business and travel expenses
- intelligence gathering for strategic and tactical planning
- distributed inventory control functions
- payroll and benefits management
- marketing, advertising and public relations
- customer support and education

It is relatively easy to imagine how any of these areas, if appropriate to a particular business intent, might benefit from a transition to digital form. These activities can all be designed to reduce the cost of doing business thus having an impact on the bottom line. Success stories abound in areas such as non-production purchasing, management of business and travel functions, and payroll and benefits. Non-production procurement accounts for 5-10% of a business's expenses. These purchases represent high overhead costs. Paper, pencils, and office chairs can represent less than 5% of accounts payable but 70% of the transactions. It is possible in an organization to distribute ordering, monitored by database triggers and aggregate them digitally for delivery to a supplier. Such a system reduces the size of the purchasing staff and allows supplies to be delivered on an as needed basis to the staff that use them reducing internal distribution cost and inventory costs. Such a system at the University of Pittsburgh reduced procurement costs by several million dollars per year. Similarly, online benefits election and direct deposit of both payroll and business expense checks can make a significant improvement in the costs of doing business.

Supply Chain Management

Where the previous section addressed primarily internal functions in an organization making use of the information infrastructure, this section addresses basic materials control in the supply chain – the movement of information and materials between organizations. The functions include:

- Production Purchasing
- Inventory Control
- Production Planning
- Logistics and Shipping

There are many books and organizations that support the development of more effective supply chain management. For the most part, supply chain management is today a *fait accompli* for the Fortune 500. It is still a challenge for small and medium sized enterprises. Given the ability to move information quickly and the desire to replace inventory with information, the basic notion is that a business wants to selectively provide information to suppliers and distributors in such a way as to insure adequate inventory on hand but no more than is needed. Ideally, a business communicates down the supply chain the information needed by the supplier in sufficient time to allow the supplier to produce and deliver the materials in a timely fashion. Given the changing nature of products that are being produced, this can become a rather complex dance. As an example, consider the types and sizes of monitors required by Dell to meet customer needs. Dell needed to communicate to its suppliers that there was a trend away from CRT monitors to LCD monitors. This required the monitor vendors to begin to make fundamental changes in what they were manufacturing. Similarly, Dell monitors how customers select memory and storage formats and other peripheral devices, and keep their vendors informed of trends in selection.

Package delivery services such as UPS, FedEx and DHL have quickly grown to provide a variety of logistics and distribution management services. For example, UPS Business Solutions includes distribution services with "more than 35 million square feet of distribution and warehousing facilities, strategically located at approximately 1000 sites in more than 120 countries."[112] Services include pick and pack order

[112] UPS (2011) UPS Business Solutions. Retrieved May 2011 from the UPS website: http://www.ups.com/bussol/?loc=en_US&viewID=browseView&WT.svl=PriNav&WT.svl=PNRO_L1

fulfillment and repair and refurbishment. They provide inventory data systems including disaster recovery, security systems, and system support and maintenance. All of these and the other business services provided by these kinds of organizations allow businesses to focus on their core competencies and use the reduction in transaction costs to make these supply chain functions more efficient.

Customer Relations Management

Maintaining customers has always been a goal of business. In this sense, customer relationship management is nothing new. Over the years good proprietors of businesses worked to satisfy their customers and knew many of their interests and preferences. It is the form and scope of the process that has changed. Today, customers are not limited by geographic location and have choices for most things they buy. Further, individual businesses have increased their customer base by one or more orders of magnitude.

Research on managing customers has increased over the last two decades. One of the things often discussed is the cost of acquiring new customers versus the cost of retaining them. What is most clear here is that the numbers vary greatly – from 5 to 8 times as much to acquire a new customer as to retain an existing one. This is not inconsistency in the research but a recognition that the cost of acquiring customers varies by industry, product, and business approach. Further, it is increasingly clear that the value of retaining customers varies over these same dimensions. That is to say. The cost of retaining a customer must be cast in the context of the value of retaining that customer. As a simple example, retaining a real estate customer who makes one purchase in forty years is less important than retaining an automotive customer who purchases a new car every two or three years and that may be less important than retaining a dining customer who eats out regularly. And that is only the first level of complication. Hotel guests might be divided between weekly business travelers and annual vacation travelers. Automobile customers might be divided between those who purchase a new luxury car every year and those who purchase an economy car every seven years. Thomas, Reinartz and Kumar present a series of equations and analysis across industries that look at acquisition and retention costs, relationship duration and profitability as the basis for making decisions about how much to spend to retain customers.[113]

[113] Thomas, Jacquelyn, Reinartz, Werner & Kumar, V. (2004) Getting the most out of all your customers. Harvard Business Review 82(7-8), pp 116-123.

Regardless of whether the business is a local restaurant or an internet giant like Amazon, the common goals in customer relationship management are to build a profile of each individual using both explicit user provided data and implicit user action data. That profile is then used to better serve the customer and more intelligently market to them. When I provide Dell with my Express Service Code, they know exactly how the machine I am calling about is configured – sales data is fed to technical support. When I order a tool from Amazon, the system is able to suggest other products people who purchased this product looked at as well as to suggest other products that are often purchased with the product. One of the interesting problems faced in customer relationship management is defining who the customer is. Consider the machines purchased from Dell. Is the customer the user of the machine, or the purchaser of the machine, or both. Situations can even get more complicated. Consider a computer program sold to some state government for the purpose of helping some department, let's say the Pennsylvania Department of Transportation, provide online vehicle registration services. Is the customer of the software vendor the government that pays for the software, the department that is serves, or the citizens that use it? These can be important questions for the modern business. Equally important is the plan for what to do with the data to increase profitability – is it up-selling or cross-selling, is it repeat business and word of mouth advertizing, etc.

Planning for E-Business

E-business clearly requires that the technology team be fully informed about the core goals of the business. Equally important, the management team needs to be fully informed about the capabilities of the technology. Management needs to understand that e-business is most often about the use of technology to do old things new ways or to do new things that have never been done before.

The infrastructure for e-business involves people, technology, and knowledge. All three legs of the stool are equally important. A lot of time and money is spent in purchasing and maintaining the technology. Today it can generally be assumed that new employees within an organization are already comfortable with PC technology. At the same time, it is still true that older employees need help in understanding how to optimally employ the new technology. It is perhaps most important to think about the knowledge base of the organization that will support the electronic enterprise. At my University, building a technical system for online registration was not anywhere near as hard as encapsulating the rules about prerequisites, co-requisites, exceptions, etc.

It is critical that the tacit knowledge within an organization be integrated in such a way as to reduce costs and prepare workers to deal with the second and third stages of e-business development. Paper based systems have to be moved to electronic form. Responsibilities for keeping electronic information stores up to date need to be developed. Care has to be taken in planning for the migration of the information from one form to the next. Finally, the duplicate paper based system needs to be eliminated.

The development of new systems assumes an informed systems analysis, design, and implementation methodology. Whether it is a traditional waterfall methodology, a rapid prototyping approach or an object-oriented methodology, the process is ultimately guided by a reengineering approach to system design. Again, the technical part is relatively easy. What is hard is that the technical staff needs to understand that the system has to have an impact on the bottom line, and the staff using the system has to take ownership of and be committed to improving the system.

New processes, channels, markets and products

The excitement of developing e-business is in projects that seek to develop new channels, markets and products. After all, for most of us this is the glamorous arena in which millionaires are made. These kinds of projects look for that new niche market or new way of doing business that will become the standard for the future. There are four types of projects here that are worth looking at in more detail:

- new processes
- new channels
- new markets
- new products

A new process can take any number of forms. Consider a set of web pages that provide customer service, or reprints of manual pages. In this case, the business is looking to provide better customer support. Rather than having customers write or call for service, the website allows customers to search a knowledge base, download patches, or replace missing documentation. This new process can also represent significant cost savings for the organization through cost reductions

A new channel takes an existing product and looks at the opportunity to sell it in a new way. Rather than ordering concert tickets by phone, a customer might order the tickets through the web. This might use a service bureau that provides credit checking and processing for a fee, or a self developed payment acquisition and checking system. Working on new channels can be tricky. For example, Avon cosmetics is carefully testing the waters of direct to consumer sales of the same products that have historically only been available through its sales force. If a self service website reduces the business for sales people or retail outlets, it may simply create a duplicate channel, and more costs, without increasing revenues.

New markets might be made possible by opening new web based channels. While the airlines open new channels for purchasing tickets online, Expedia.com and Bookit.com have created new markets for bundled vacation services including flights, car rentals, hotel reservations and activities. Similarly, autotrader.com, vehix.com and carmax.com, to name just a few of the competitors, have created markets for aggregated services for finding, comparing, vetting (via carfax.com), transporting, and financing a used car.

New products range from truly new products such as apps for smart phones to derivative information products developed within organizations doing other form of business. For example, a bank might provide a currency conversion service. A software house might provide a new kind of plug-in for browsers. A symphony might provide pay per listen live broadcasts of concerts.

Some Concluding Observations about Business

The discussion in this chapter has been very conceptual and general. It has not included lengthy discussions of service oriented architectures or enterprise resource planning (ERP) systems. Rather it has addressed the conceptual reasons why these systems are being put in place. What has been discussed is equally true whether you are part of General Electric involved in the development of a massive ERP system or the owner of a local gas station maintaining inventory. The same principles used by Amazon to develop a customer relations management system can be applied inexpensively by a local restaurant. Sending a reminder to good customers about holiday event specials at a restaurant is the same as cross selling Amazon customers.

A Few Cautions

E-business reflects a transformation of business to rely on digital information with all of its cost savings and ease of manipulation. In the transition to digital information, organizations need to be appropriately conservative to make sure the transition takes place as planned. Some people in an organization will resist the move and continue to prefer more traditional forms of doing business. Once the transition has been made, the duplicate infrastructures need to be removed. If documents in all the forms they exist in organizations from memos and letters to product manuals to annual reports to invoices absorb six or seven percent of the organizational expenditures, the physical infrastructure must be discarded. While it didn't take long for typewriters to be abandoned, file cabinets and printers have been more resilient. Besides the cost of the items, they occupy expensive space and require human effort to maintain. The duplicate systems represent a drain on profits and must be aggressively rooted out.

A corollary of this caution relates to document management. One of the most important characteristics of digital information is the ease with which it can be changed. Organizations have historically archived important documents. As just one example, colleges maintain bulletins which describe the courses they offer and the requirements for various degree programs. If someone wanted to check on the degree requirements and course descriptions for a degree in psychology from the College of the Holy Cross in 1972, it would simply be a matter of accessing the college catalog for that period. As these documents have gone on the web, it has been possible for institutions to make minor editorial changes in real time to correct errors and this is good. However, it is not unusual to find situations in which major revisions of substance get made with no record of the change. This can sometimes have significant negative consequences in the long term. Organizations need to carefully transition responsibilities for document and information management in digital form. This includes both change management policy – who initiates and who approves changes, version control, and responsibility for archiving and disposal.

About once a week, there is a news story about some important security breach where personal information is exposed or some financial catastrophe occurs. Digital information presents security concerns given the ability to access it from a distance and steal massive amounts of it in a matter of seconds. These security concerns are real and need to be addressed. What organizations, especially smaller organization and those that are not intensively involved with financial data, often forget is that the largest security threat is insider action. “Joe would never misuse information or sell it on the outside.” Unfortunately, time and again, we find that Joe is involved much to the surprise of the boss. The other security concern relates to appropriate backup. I hear more complaints about data loss than almost anything else, especially from small organizations. The cost of offsite storage just doesn't seem worth it – or not worth doing at regular periods. The unanticipated power outage can destroy a full day's work for a small business.

The Next Big Thing

Predicting the next big thing in e-business is a difficult task, and depends on your perspective. Investments might be directed toward network or hardware infrastructure, service providers, general e-businesses, or e-tailers. I have always found basic infrastructure such as UPS and Cisco very attractive. Indeed, these firms do very well supplying the infrastructure that needs to be built up regardless of the companies that win or lose at the commerce level. As indicated earlier, UPS, in a fashion similar to Amazon, has leveraged and expanded its business to fill the many niche services required by e-business. Also attractive are companies that seek to make the physical interface to digital information. No company in recent history has been more successful than Apple with the introduction first of the iPod, then the iPhone and most recently the iPad. These technologies work at the human level and make the vast world of digital information more accessible. While these areas of business will continue to grow, they are not what most people are looking for when they think about the next big thing.

In 1994, Jeff Bezos quit his job as an investment banker and began to think about what might effectively be sold over the internet. He observed several things about the book selling industry. First, there are more than 3 million books in print at any given point in time, and bookstores can hold only a fraction of them. Second, even the biggest operations such as Barnes and Noble accounted for only a small share of the market. Third, books are products that don't require the consumer to touch them to make a purchase decision. Fourth, the large number of publishers and the existence of big distributors made inventory management fairly easy. Finally, they were well suited to parcel delivery. Amazon lost money for several years as it pioneered the software development that would later make them famous for easing payments, profiling customers, and managing customer satisfaction. Now they have managed to make that software suite a secondary business assuring their prominence in the online retail world. In contrast, eBay was founded more or less by accident. Pierre Omidyar, a computer programmer, added a piece to his personal website called "AuctionWeb." When he found in 1995 that he was able to sell a broken laser pointer to an individual who collected broken laser pointers, he realized that there could be a niche for selling collectibles. Over the succeeding years, eBay expanded beyond collectibles and introduced new mechanisms for payments and trust management in an online environment.

The history of the web is intimately tied to efforts to overcome its Achilles' heel – finding the information you want. The basic problem is that link traversal is not an effective method for finding the information you want. While the vision may scale to an individual's or maybe even an organizations personal information store, it does not fare well at a global level. Historically, the world's knowledge has been classified by librarians and that is how things started on the web. In 1994, two graduate students at Stanford, Jerry Yang and David Filo created a website called "David and

Jerry's Guide to the World Wide Web." It was a collection of websites they had visited organized in a hierarchy, or classification system. As the Web grew exponentially, human classification did not scale very well. Computer power was on the rise and researchers at DEC's Western Research Laboratory were exploring new software programs that could collect web pages and new mechanisms for creating full text databases. That led to the development of the AltaVista search engine that could search for the occurrence of particular words across a vast index of web pages. Without going into the details of relevance ranking, AltaVista used traditional techniques for deciding which of 1000's of pages identified as containing certain terms might be the ones that would answer an individual's query.

In 1997, Larry Page and Sergey Brin, graduate students at Stanford founded Google. They had been working on digital library research. For hundreds of years, scholars have used citation analysis to establish both the importance of a given publication and to understand how it relates to other work. Put most simply, if there are two articles on some subject, and one of them is cited by 1000 scholars and the other is cited by only one or two, the first article is like to be more important. On the web, a search engine could find all the pages that contained a given term, and using an algorithm, called "Term Frequency-Inverse Document Frequency" it could rank which documents were likely to be more relevant. Basically, if a document has a lot of words in it that match the one being searched for, and if that term does not occur in a lot of documents, it is likely to be more relevant. Brin and Page developed an algorithm similar to the citation index which said, in simplified form, if a page is pointed to by a lot of pages and if the pages that point to it are themselves important, the page is likely to more relevant. The long and short of the story is simple, people found this form of ranking better than more traditional techniques, but that is just the beginning of the story.

What makes Google the giant it is was the development of a financing model. It is as simple and as elegant as the relevance formula. While the actual financing model is as complex and rich as the full ranking model, at the core it is elegantly simple. It begins with the assumption that when someone enters terms for a search they might be looking for something to buy. It also assumes that there are businesses that might want to be exposed to these potential consumers and that they would value that exposure differentially. Thus, a dog food vendor would not be willing to spend as much to be made visible to a potential consumer as a jeweler would. When individuals search on Google, the top section of the results page as well as the panel on the right, labeled as "Ads" display links that identify businesses that have paid to be made visible to the searcher. When the searcher chooses to follow one of those links the advertiser is charged for delivering that potential consumer to the vendor's site. It is this revenue stream that has funded the development of all of the various Google enterprises each of which has different and complimentary revenue models.

As we have already described, social networking sites began to emerge in 2002 first with Friendster, then MySpace and finally in 2004 Facebook, which has emerged as the most popular site. These sites are not directly engaged in sales or information finding, but rather focus on communication and community. Facebook combines the click-through revenue model of Google with the profiling techniques of Amazon. Importantly, Facebook has established a new sense of place where visitors stay within the space for greater periods of time.

These developments will all continue. In the era of immediacy, it is likely that the next big development will emerge out of these significant trends. I believe that there are three possible seeds that will generate another significant surge. The first is DWIM – Do What I Mean. The profiling and collaborative intelligence developments make it likely that at some level we will begin to develop anticipatory analyses that will allow systems to anticipate our needs. Systems will find a vacation trip I really want to take based on what it knows about me and let me know. This is one form of just-in-time(JIT) information. Other forms of JIT information will be provided by the ubiquitous information environment. Rather than have to look something up in advance, a system will emerge that will provide it to us exactly when we need it. This kind of information is made more possible by location awareness. As cell tower and GPS tracking systems evolve, there will be an increased delivery of information based not only on my profile but on where I am and what I am doing.

Chapter XII: Final Thoughts

Waves

In **The Third Wave**, Alvin Toffler outlined three massive culture shocks or waves of change. The first was the transition from hunter gatherer society to an agricultural society where food was cultivated and controlled by human endeavor. The second wave was development of industrial society and the mass production of commodities. The third wave was the technological wave variously called the post industrial society or the information age. As we have discussed, the revolution combining the computer, networking, and digital information is and will continue to allow levels of customization and individualization in the production of materials. It has also resulted in new forms of communication.

One of the most interesting impacts of these advances is the unpredictable and profound nature of the changes that occur in the aftershocks. These come in two forms. By way of example, the industrial revolution was bases on the application of power technology. That included water power, steam power, the internal combustion engine, and the electric motor. When we envision the beginning of the industrial revolution, it often includes the vision of steam engines, but of course it had echoes that involve other power technology. Many of the important devices we use today are dependent on internal combustion or electricity. (Steam powered hedge clipper, lawn mowers, televisions, or radios present a conceptual challenge.)

The second and more important kind of aftershock involves the secondary and tertiary impacts of the revolution. Consider as an example the growth of factories and businesses in places where transportation allowed for the movement of raw materials and finished products. The growth of cities can be defined as a secondary impact of the industrial revolution. The industrial revolution also had an impact on social stratification adding a significant middle class. Unionization, specialization and alienation may be viewed as tertiary impacts of the industrial revolution. Looking at a more specific technology such as automobiles can be tied to a number of secondary and tertiary impacts such as the growth of the suburbs, mass construction of roads and parking lots, changing family co-location, compressing space and time, climate changes due to the proliferation of roads and parking lots, changes in childhood morbidity, etc. What will the secondary and tertiary impacts of the digital information revolution be? Trying to predict these impacts is dangerous, but fun. Two aspects of the current revolution that appear to have some potential to induce secondary change are virtualism and telepresence.

Virtualism is likely to take many forms. It may be virtual organizations, virtual social groups, etc. At the extreme, one might imagine organizations made up of ad hoc sets of individuals gathered in a form of virtual organization. Today, organizations need to maintain employees with specialized talents as whole units, whether or not there talents are needed on a continuous basis. In future organizations, it may be more cost-effective to retain access to these individuals who might share their talents across multiple organizations. The model might be something like the accounting model where the relationships are fairly stable, but purchased on a service basis rather than employed full time. These would be more than temp jobs. They would represent an extreme form of outsourcing, but with the intent of long term variable intensity relationships. Given these kinds of relationships, it may be less important for people to live near where they work. There may be more need for temporary residences. Where it makes sense the meetings may move to virtual forms.

Virtual organizations may make use of telecommunications to run meetings that provide an acceptable substitute for face to face meetings. At some levels, major sports events are achieving a level of virtual presence. Colleges and universities are providing access to lectures online that provide for new forms of educational opportunities. We also pay differential amounts for different levels of presence. One thousand dollars or so will buy travel, a room in New York and tickets to the original Broadway cast of some show. For about a tenth of that, I could buy tickets for the traveling cast of that show. For a tenth of that, I could go to a movie theater and share the experience of a telepresence view of that show with other movie goers, and finally for a tenth of that – about a dollar, I could rent the DVD from Netflix streamed to my home for personal viewing. This technique may allow for new forms of presence in many other activities. While I am unlikely to want a virtual vacation in Playa Del Carmen from chilly Pittsburgh, there are other trips that might benefit from telepresence. For example, a few years back I made a side trip to Mainz while I was in Germany. My goal was to visit the place where Gutenberg worked with Schöffer to develop moveable type printing. One of my bucket-list goals is to travel to all the places where significant events in information processing took place. It might be the case that it would be more effective to take some of those trips via telepresence. Indeed some enhanced form of telepresence might be better than face to face. Imagine a wrap around experience – maybe available in major cities by reservation that would allow robotic control of devices that would provide a personal experience and an experience that might not be possible in person. Without worrying too much about the details, it is not hard to imagine a personal video tour of the Sistine Chapel that allowed views I could not obtain face to face. It might be possible to fly over a volcano or trek through the catacombs of Paris.

Memex and the Global Brain

Language and particularly the written word may be viewed, from an information science perspective, as a mechanism to supplement to human memory. Spoken language limited humans to being able to store and transmit information that could be accumulated and transferred from human memory to human memory. It was subject to significant transmission errors and storage errors. Writing allowed a quantum leap in our ability to extend memory and store information with fewer errors in storage – although it may be that for some things written language is subject to more errors in transmission related to the translations required. Printing allowed a quantum leap in our ability to disseminate information. Telecommunications and computers provide a quantum leap in the ability to capture, storage, and dissemination both symbolic messages as well as non symbolic sensory data.

The question that one might ask at this point is how a "digital information fabric" might impact our ability to accumulate, manipulate, record, and disseminate information. There are indications that digital information is not only supplementing memory but thinking as well. Thus, very early written records simply codified facts – Ahmed has 243 camels. It was not long before processes were also written down. The height of a tree may be calculated by taking a 3/4/5 right triangle and aligning the base with the base of the tree and the hypotenuse with the top of the tree. The distance from the corner of the triangle to the base of the tree will be 3/4^{ths} of the height of the tree. What is happening now with computers and information is not only a reflection of the factual record, but the thinking process itself being encoded.

Personal Collective Brain

What would constitute a personal collective brain? How could we take the accumulated knowledge of a group of individuals working on a set of closely related problems and make them accessible over space and time to someone who could benefit from the knowledge. Consider a soldier on the battle field who might access the collective knowledge of military personnel who have fought in situations similar to the one the soldier is now in. Generals from Julius Caesar to Erwin Rommel and Norman Schwarzkopf have provided commentary and analysis of Battle field efforts. In the 1970, the US military began to aggressively record after action reviews (AARs) that provided a mechanism for analysis of what was intended, what occurred, why were the actual outcomes different from the planned outcomes and what was learned.[114] We are not far away from finding better ways to organize structure and use this data. Take it one crazy step further. Imagine that with more substance the final system might be like Obi-Wan Kenobi from Star Wars saying to Luke Skywalker – "trust the force Luke." More realistically, it would seem reasonable to imagine that the field commander being fed digested AARs might suggest to a field soldier that he should "watch out for insurgents behind the door of the Office building on the right." The idea would be that the system identifies the relevant information and brings it to bear at the right time in the process. Assistants of this type are being built today for various purposes, and AARs have moved beyond the military and into organizations that are conducting multi level analyses of their engagements.

Information Fabrics

What would an information fabric look like? Imagine that everything that is observed about the world is captured as it is observed – from the trivial to the profound. It is all incorporated into an information fabric in the ether. Similar observations are reinforced and noise is eliminated. Users continuously receive subtle feedback. New information, it is moved to the appropriate place in the fabric, cataloged, and made accessible as appropriate to the various communities it may impact. Kind of like the collective brain, but incorporating all of human knowledge. Nothing is ever lost, but at the same time, the fabric is self correcting. That is, bad or erroneous information is noted and marked off. New information is highlighted and shown. Redundant information is simple overlayed on the existing store. People can "rediscover" things but not record them.

A Personal Agenda

It occurs to me that one question I might have on reading this book is how it will make my personal life different. Of course that depends on

[114] US Army (1993) A Leader's Guide To After-Action Reviews. Training Circular 25-20, Department of the Army Washington, DC, 30 September 1993. Retrieved May 2011 from the Air University website: http://www.au.af.mil/au/awc/awcgate/army/tc_25-20/tc25-20.pdf

how old you are and what you do. Thinking of the grandchildren I don't yet have, I have begun the process of organizing digital forms of all the recordings I have of my children. It is also for my children. Twenty years of slides, negatives, prints, super 8mm film, VHS tapes and super VHS took almost five years to put in digital form. The last ten years of pictures were already in digital form as are the last five years of video. One of the very interesting side effects of this process was the ability, after the video was converted to digital form was the ability to create DVD's for my family members that included all of the videos that included visits we had had over the years and the growth of our children.

Knowing that with the passing of both of my parents there was some loss of important photo's and memorabilia that went to one or another of the siblings, it is now possible for both of my children to have a complete photographic record of their upbringing to share with their children. I regret that many of the artifacts that existed in paper form have been lost – grade reports, projects, papers are mostly gone with the exceptions of those we explicitly put away for them. If my children were born today, I would make every effort to scan important papers and documents in their lives that will continue, at least for a few years to exist in paper form. They will want at some time, probably not until they themselves have children, to look back on key events. It is all too easy to lose track of important events that have not been recorded in some form. I would also look to see if I could find a way, or encourage them, to keep as much digital information as they can. I have every email that I have exchanged on important projects in my professional life and with significant others. Whether I will ever go back to look at those or think about them is not clear, but I will have that choice.

Interestingly, the status of my publications, which I view as one of the two foci of my life's work, is less clear. Early publications were not made available in digital form, and are not likely to be. Fortunately the tenure process encouraged me to keep physical copies of most publications and at least one copy of each book, but at some point I will have to scan those – at least as images. The second focus of my work is teaching, which I consider to be my greatest talent, but in general that is not available to my children. I have made a couple video lectures, but they are difficult to capture and are not the best indication of my teaching. I need to work on seeing if I can find ways to better capture some of those pieces. Over the years, I have written more than 100 pieces of poetry and while there are digital copies of the more recent efforts, collections from 1976 and 1982 exist only in paper form and of them I can find only a few printed copies. My next goal is converting those early poems into digital form and combining and organizing all of the work into some cohesive form.

Chapter XIII: Selected Readings

While I like to believe that some of the thoughts in this book are brand new, I suspect that most of them have roots in things I have read over the last fifty years. I make an effort to try to understand the history of anything I teach. I find that often the original vision is incredible simple and clear. With time, things get more complex and hard to understand. I have tried to identify several books and articles that have made a contribution to my own thinking. Some are closely associated with this book, some provide important historical insights, and others are somewhat tangential but so interesting as to merit a recommendation if you want to learn more about the basis for this book.

Important Insights

Vannevar Bush, "As We May Think", Atlantic Monthly, July, 1945, pp 101-108. (http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/)

Many attribute the origin of hypertext and ultimately the World Wide Web to this article by Vannevar Bush. There is an interesting follow-up article that appeared in Life magazine a couple months later. (Bush, V. As We May Think. Life Magazine 19, 11 (September 1945), 112-114, 116, 121, 123-124.) The Life magazine article shows some artist renderings of the various devices Bush was envisioning.

Douglas Engelbart, Augmenting Human Intellect: A Conceptual Framework (http://www.dougengelbart.org/pubs/augment-3906.html)

This is not an article for a casual reader, but it provides great insight into the man who made quantum leaps in digital information. Douglas Engelbart was decades ahead of everyone

else at a time when being years ahead was significant. He has now begun to receive the recognition he deserves.

J. C. R. Licklider, Man-Computer Symbiosis IRE Transactions on Human Factors in Electronics,Volume 1, pages 4-11, March 1960 http://groups.csail.mit.edu/medg/people/psz/Licklider.html

*It is said that Licklider, as head of the Information Processing Techniques Office (IPTO) at the Department of Defense Advanced Research Projects Agency in 1962 controlled more money that could be devoted to advanced computing concepts than the rest of the US government. His 1960 article on Man-Computer Symbiosis presented an articulate view of what would emerge over the coming decades. He was succeeded at IPTO by Ivan Sutherland and later Robert Taylor who would finally begin to build what would become the internet. Licklider is credited as the father of the internet, appropriately I believe, but Taylor also deserves credit. A joint article by Taylor and Licklider may also be of interest. (*J.C.R. Licklider and Robert W. Taylor The Computer as a Communication Device. In: In Memoriam: J. C. R. Licklider 1915-1990. *This memorial publication includes both the original article by JCR licklider and a this companion piece by Licklider and Taylor. http://memex.org/licklider.pdf)*

Shoshana Zuboff, In the Age of the Smart Machine: the Future of Work and Power. New York: Basic Books, 1988.

Zuboff presents one of the most coherent discussions of the application of computer technology to work. Given the time at which she was writing, the views she expresses are amazing. A fun and illuminating read. It shows the impact of information without all the glitz of technology to distract.

Nicholas Negroponte, Being digital. New York: Knopf, 1995.

Negroponte is one of the great visionaries of our day. His primary message was to move bits not atoms. Being Digital is a small little book, his only I believe, that presents a human and understandable view of what is going on today. He was co-founder of the Media Lab at MIT with Jerome Wiesner. He has also been a prime mover in the "One Laptop per Child" project.

Robert Lucky. Silicon dreams: information, man, and machine. New York : St. Martin's Press, 1991.

Regrettably, this book is out of print, but still available in used form from several sources. Lucky will be known to professionals for his regular column in IEEE Spectrum. Many of those columns became the basis of various chapters in the book and more recent columns can be accessed online (http://www.boblucky.com/spectrum.htm). A senior researcher at Bell Labs and Bellcore, Lucky has a unique ability to make complex ideas clear to the thinking person. My vision for this book was in large part inspired by Lucky's writing.

Michael L. Dertouzos, What will be : how the new world of information will change our lives. San Francisco, Calif.: HarperEdge, 1997.

Michael L. Dertouzos, The unfinished revolution : human-centered computers and what they can do for us. New York: HarperCollins, 2001.

Michael Dertouzos directed the Laboratory for Computer Science at MIT for more than 25 years. A broad ranging researcher at the cutting edge of many projects, he was able, perhaps more than any other scientist to speak with authority about the trajectory of the information revolution. What will be is a readable vision of the future informed by an encyclopedic recollection of what has been. His last book, published posthumously, provides a technologist roadmap for what we need to do, written in a way that almost anyone can understand.

Peter J. Denning & Robert M. Metcalfe, (Eds.) Beyond Calculation: The Next Fifty Years of Computing. Springer, 1998

This insightful collection of essays describes what is beyond calculation – communications. Published for the fiftieth anniversary of computing in 199, the essays focus on the next fifty years. The contributors see a future more about the computer as communication device than as calculation device. Some of the more interesting essays address such topics as the end of the nation state, information warfare and artificial intelligence.

Historical Perspective

Elizabeth L. Eisenstein, The printing press as an agent of change: communications and cultural transformations in early-modern Europe. Cambridge University Press, 1979

The printing press was nominated by several organizations as the most important invention of the second millennium and appeared on most top ten lists. Given the clock, the steam engine, the transistor, vaccines, the computer, etc. it is clearly an impressive development. Eisenstein's two volumes tell the story eloquently and provide a compelling explanation of the impacts of mass production printing.

Tim Berners-Lee, Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor. San Francisco: Harper, 1999.

*It is fun to have the opportunity to view the early development of the World Wide Web through the eyes of the man at the center. Most importantly, Berners-Lee shares his vision of the Web as a space for collaboration – a part of the dream that is only beginning to appear now. In the book, Berners-Lee opens up the topic of the semantic web which is described in more detail in an article published in Scientific American (*Berners-Lee, T., Hendler, J., Lassila, O. 2001. The Semantic Web. Scientific American. 294(5): 28-37.*) Unfortunately, there is no layman description of the semantic web although much of the work is publicly available through the World Wide Web Consortium – (*http://www.w3.org/standards/semanticweb/*)*

Michael A. Hiltzik. Dealers of Lightning: Xerox PARC and the Dawn of the Computer Age. Harper Business, 2000.

Hiltzik tells the story of the people who developed the technology at Xerox PARC. As he reports, there was a time when many considered the researchers at PARC to be some of the brightest in the country – they say 40 of the top 100 computer scientists were at PARC. The book reads like a suspense novel moving from building Bose speaker knockoffs to exposing the revolutionary technology of the Star to Steve Jobs and his engineers at Apple. A more sobering view of PARC focusing on the corporate failure to capitalize on the inventions may be found in: Douglas K. Smith and Robert C. Alexander. Fumbling

the Future. HarperCollins Publishers, 1989. Tells the business story of Xerox PARC.

Special Focus

John Seely Brown, Paul Duguid The Social Life of Information, by Harvard Business School Press, 2000

John Seely Brown is another PARC alumnus – former chief scientist and Director. With Paul Diguid, this book looks at information in organizations and the changing nature of documents. There discussion of "sticky" and "leaky" information is particularly insightful

Mark Buchanan, Nexus: Small Worlds and the Groundbreaking Theory of Networks. Norton, W. W. & Company, Inc. 2003. ISBN 0-393-32442-7.

*Mark Buchanan builds out the story of small world networks which arise from a small, but important article in nature by Ducan Watts and Steven Storgatz. (*Watts, Duncan J.; Strogatz, Steven H. (June 1998). "Collective dynamics of 'small-world' networks." Nature 393 (6684): 440–442. doi:10.1038/30918. PMID 9623998.*) Until the introduction of small world network theory, graph theory had been absorbed with the study of regular and random networks. Small world networks have very special characteristics. Since the "discovery" (Watts and Storgatz developed a mathematical description of the phenomenon), Small World Networks have been the basis of a significant amount of research in a wide variety of fields. (The internet and the World Wide Web are both small world networks.)*

Ravi Kalakota & Marcia Robinson, e-Business 2.0: Roadmap for Success (2nd Edition) Addison-Wesley Professional; 2 edition (December 11, 2000)

After a decade, Kalakota and Robinson's treatment of e-Business is still the best conceptual description of what it is all about. The discussion in this book is enriched and informed by their views.

Donald Norman, The Psychology of Everyday Things. New York, NY, US: Basic Books. (1988).

Donald Norman has written a number of books on the human interface to computer systems and other devices. His later works are solid and well thought out, but this first book is both illuminating and approachable.

Ray Kurzweil, The age of spiritual machines: when computers exceed human intelligence Penguin, 2000

Kurzweil is famous for his prediction of "the singularity" – the ultimate human machine symbiosis. As an inventor who solved very difficult problems – intelligent text recognition and speech recognition, his perspectives are worth considering. Whether or not you buy the view of a transhuman future, it is hard to dismiss his views on a future where computational systems will grow to exhibit what we consider intelligent behavior.

George B. Leonard, Education and Ecstasy. New York, Dell Publishing Delta Books, 1968.

Of all the books that have stayed in my personal library related to this subject, Leonard's is the oldest. His vision of the future of education is fueled by the theories of Dewey, Bruner, and Piaget and infused with the humanist perspectives of Goodman, Holt and Kozol. He blends the perspective of a technologist who imagines an electric car arriving at a sustainable school where children are intimately tied to patient intelligent collaborative computer systems dedicated to personalized holistic education.